Spring 4-23-2020

Finite Population Corrections for Two-Level Hierarchical Linear Models with Binary Predictors

Steven Svoboda
University of Nebraska - Lincoln, ssvoboda@huskers.unl.edu

FINITE POPULATION CORRECTIONS FOR TWO-LEVEL HIERARCHICAL LINEAR MODELS WITH BINARY PREDICTORS

by

Steven J. Svoboda

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Psychological Studies in Education

(Quantitative, Qualitative, and Psychometric Methods)

Under the Supervision of Professor R.J. de Ayala

Lincoln, Nebraska

May, 2020

FINITE POPULATION CORRECTIONS FOR TWO-LEVEL HIERARCHICAL LINEAR MODELS WITH BINARY PREDICTORS

Steven J. Svoboda, Ph.D.

University of Nebraska, 2020

Adviser: R.J. de Ayala

Answering social science research questions about clustered data necessitates collecting data using sampling schemes, which may result in hierarchical data structures. Hierarchical liner modeling (HLM) techniques are required to account for the interdependency of observations due to clustering. However, traditional HLM assumes the target population is infinitely large or near enough to infinitely large for practical purposes (i.e., the sample consists of less than 5% of the target population). Often times, the assumption of an infinitely large target population may not hold.

The current study was conducted in two separate phases using Monte Carlo simulation methods. First, the continuous predictors study evaluated a finite population correction (FPC) method for a few number of large clusters. The degree of relative bias in unadjusted standard error estimates exceeded .05 and was non-ignorable when the number of clusters sampled was greater than 20. The finite population correction adjusted standard error estimates exhibited acceptable levels of relative bias across most simulation conditions. However, finite population correction adjusted standard error estimates were negatively biased when the number of clusters sampled was few (i.e., 20 clusters). The continuous predictors study also examined standard error estimates from a finite population bootstrapping alternative. The finite population bootstrap estimates did

not perform well and severely underestimated the empirical standard errors across all conditions.

Second, the binary predictor study evaluated the efficiency of the finite population correction method for a level-2 binary predictor. Standard errors for a balanced binary predictor (i.e., binary predictors with a relatively constant 50:50 prevalence between groups) functioned similarly in terms of bias as continuous predictors. The relative bias in the finite population correction adjusted standard errors for a balanced predictor was smaller than the relative bias in unadjusted standard errors when at least 30 clusters were sampled. For a discrepant or unbalanced binary predictor (i.e., 20:80 prevalence), finite population correction adjusted standard errors were only acceptable when 60 clusters were sampled.

The current study demonstrates the need for applied researchers to explicitly state their target populations, examine their sampling fraction, and consider the FPC adjustment. Doing so yields more accurate inferences for finite populations.

Dedication

For Hadley & Huxley

Acknowledgements

I am *infinitely* grateful for the assistance I received during my graduate studies. Thank you to my doctoral committee, particularly Ralph, for your support and valuable feedback.

Dr. Mark Lai, thank you for piquing my interest in finite populations and for your correspondence over the past months. I hope you learn something interesting from my dissertation.

This work was completed utilizing the Holland Computing Center of the University of Nebraska, which receives support from the Nebraska Research Initiative. Specifically, thank you Carrie Brown. Without your tips, my simulation might still be running.

Thank you to all the child care providers (also known as family members) including but not limited to: Grandpa Jim, Grandma Jan, Great-Grandma Kathy, Great-Aunt Maryann, *Dad's* Aunt Megan and Aunt Frank, and *Mom's* Aunt Megan & Uncle William. And thank you to my wife, for the luxury of having a partner with so much methodological expertise.

Finally, thank you Carolee. A lot of people deserve credit for aiding me throughout this process, but none more so than you. I would have never made it this far without your help.

What a long, strange trip it's been.

– The Grateful Dead

Table of Contents

List of Tables ix

List of Figures x

CHAPTER I. INTRODUCTION 1
- Purpose 5
- Significance 7

CHAPTER II. LITERATURE REVIEW 8
- Sampling 9
  - Probability & Simple Random Sampling 9
- Modes of Statistical Inference 13
  - Competing Frameworks 13
  - Integrated Framework 25
- Why use Finite Population Corrections? 33
  - Target of Inference and Fixed vs. Random Effects 33
  - Multilevel Studies in which the Population is Finite 36
- Finite Population Corrections 38
  - Estimating Finite Population Parameters 39
  - FPC for the General Two-Level Linear Mixed Model 42
- Bootstrapping 48
  - Multilevel Bootstrapping 51
  - Finite Population Bootstrapping 52
- Summary 53
- Current Study 54

CHAPTER III. METHODS AND PROCEDURES 56
- Data Generation 56
  - Continuous Predictors Study 56
  - Binary Predictor Study 58
  - Generating Data Using Copulas 58
- Simulation Conditions & Their Justification 59
  - Continuous Predictors Study 59
  - Unique Conditions for Binary Predictor Study 60
  - Constants 61
- Procedure 62
  - Computational Intensity Pilot Study 63
- Evaluation Criteria 64
  - Relative Bias 64
  - Mean Square Error 65
  - Root Mean Square Error 65
  - Binary Predictor Coverage 65

CHAPTER IV. RESULTS 67
- Continuous Predictors Study 67
  - Level-2 Effects 67
  - Level-1 Effect 71
- Binary Predictor Study 73

Continuous Predictors' Effects 75
Binary Predictor's Effect 75

CHAPTER V. DISCUSSION 79
Main Findings 79
RQ1a & b (Comparing SE_0s to SE^{FP}s for a Few Number of Large Clusters) 80
RQ2 (Comparing SE^{FP}s to SE^{FPboot}s) 82
RQ3 (Comparing SE_0s to SE^{FP}s for Binary Predictor) 82
RQ4 (Comparing SE^{FP}s to SE^{FPboot}s for Binary Predictor) 84
Limitations & Future Directions 84
Implications for Applied Research 87

References 90
APPENDIX A: ABBREVIATIONS & NOTATION 98
APPENDIX B: R CODE USED FOR SIMULATION 99
APPENDIX C: CONTINUOUS PREDICTORS' RESULTS FROM THE BINARY PREDICTOR STUDY 108

List of Tables

Table 1.1. Summary of Design Factors for the Continuous Predictors Study 6

Table 1.2. Summary of Design Factors for the Binary Predictor Study 6

Table 2.1. Properties of Probability Sampling and Alternatives to SRS 10

Table 2.2. Infinite and Finite Population Equations 39

Table 2.3. Summary of Design Factors from Lai et al. (2018) 45

Table 4.1. Average Relative Bias Across Conditions for Continuous Predictors Study ... 67

Table 4.2. ANOVA for Relative Bias in $\boldsymbol{SE_0}$ of $\boldsymbol{\gamma_{01}}$ for Continuous Predictors Study ... 67

Table 4.3. Mean Square Error and Root Mean Square Error for $\boldsymbol{\gamma_{01}}$ in Continuous Predictors Study 69

Table 4.4. ANOVA for Relative Bias in $\boldsymbol{SE_0}$ of $\boldsymbol{\gamma_{02}}$ for Continuous Predictors Study ... 70

Table 4.5. Mean Square Error and Root Mean Square Error for $\boldsymbol{\gamma_{02}}$ in Continuous Predictors Study 71

Table 4.6. ANOVA for Relative Bias in $\boldsymbol{SE_0}$ of $\boldsymbol{\gamma_{10}}$ for Continuous Predictors Study ... 71

Table 4.7. Mean Square Error and Root Mean Square Error for $\boldsymbol{\gamma_{10}}$ in Continuous Predictors 73

Table 4.8. Number of Populations in Binary Predictor Study with Complete Sample Replications 73

Table 4.9. Number of Populations in Binary Predictor Study with Complete Bootstrap Replications 74

Table 4.10. ANOVA for Relative Bias in $\boldsymbol{SE_0}$ of $\boldsymbol{\gamma_{02}}$ for Binary Predictor Study 75

Table 4.11. Average Relative Bias in *SEs* for $\boldsymbol{\gamma_{02}}$ in the Binary Predictor Study 77

Table 4.12. Mean Square Error and Root Mean Square Error for $\boldsymbol{\gamma_{02}}$ in Binary Predictor Study 77

Table 4.13. Proportion of Populations with Interval Estimates Including $\boldsymbol{\gamma}$ and 0 78

Table C.1. Mean Square Error and Root Mean Square Error for $\boldsymbol{\gamma_{01}}$ in Binary Predictor Study 108

Table C.2. Mean Square Error and Root Mean Square Error for $\boldsymbol{\gamma_{10}}$ in Binary Predictor Study 109

List of Figures

Figure 3.1. Path diagram for data-generating model 57
Figure 4.1. Percentage relative bias in *SE*s for $\boldsymbol{\gamma_{01}}$ in the continuous predictors study ... 68
Figure 4.2. Percentage relative bias in *SE*s for $\boldsymbol{\gamma_{02}}$ in the continuous predictors study ... 70
Figure 4.3. Percentage relative bias in *SE*s for $\boldsymbol{\gamma_{10}}$ in the continuous predictors study ... 72
Figure 4.4. Percentage relative bias in *SE*s for $\boldsymbol{\gamma_{02}}$ in the binary predictor study 76
Figure C.1. Percentage relative bias in *SE*s for $\boldsymbol{\gamma_{01}}$ in the binary predictor study 108
Figure C.2. Percentage relative bias in *SE*s for $\boldsymbol{\gamma_{10}}$ in the binary predictor study 109

CHAPTER I. INTRODUCTION

Physical, behavioral, and psychological research questions often relate to hierarchical or multilevel data systems (Mass & Hox, 2004). Answering many of those social science research questions necessitates collecting data using sampling schemes other than traditional simple random sampling (SRS), such as cluster or multistage sampling, resulting in hierarchical data structures (Lai, Kwok, Hsiao, & Cao, 2018). Refer to Appendix A for a list of abbreviations and notation used throughout this manuscript. Examples of hierarchical data structures include, but are not limited to students nested within classrooms or schools, employees nested within supervisors, and patients nested within hospital wings. Standard, single-level regression techniques are not appropriate for modeling these data structures, even if the analysis includes only level-1 predictors, because failing to account for hierarchical data structures likely violates the assumption that errors are independent of each other and identically distributed. This violation results in biased standard errors associated with the regression coefficients, which in turn, leads to increased Type I error rates and erroneous interpretations of statistical tests (Mass & Hox, 2004; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012).

Hierarchical data structures exist in nature whether or not psychologists and behavioral scientists recognize their existence and account for the nesting of their subjects within higher order units in applied research. Unlike standard single-level modeling, hierarchical linear modeling (HLM) techniques account for nested data structures and have received considerable attention in recent decades because of the realization that empirical research in the social sciences often deals with data, which are

hierarchical in nature (Lai et al., 2018; Maas & Hox, 2004; Raykov, 2010; Raykov, Patelis, Marcoulides, & Lee, 2016). However, the theories used to develop HLM assume observations are sampled from an infinitely large population with "little attention given to situations where, at some higher level, the sampled units are a subset of a finite target population" (Lai et al., 2018, p. 94). Meeting the assumption that observations are drawn from an infinitely large population may not be tenable in many applied settings and failing to meet this assumption necessitates the use of a finite population correction.

Cochran (1977) introduced finite population corrections (FPCs) based on the sampling fraction $f = n/N$ where n is the sample size and N is the finite population size. In practice, FPCs may be ignored if f does not exceed 5%. Larger samples relative to their populations require FPCs because ignoring large sampling fractions results in biased standard errors (Cochran, 1977). Applied researchers should identify their target populations, examine their sampling fraction, and consider using FPCs because applying FPCs yields more accurate inferences for finite populations.

FPCs are not common for single-level regression models even though using them may yield more accurate inferences (Lai et al., 2018). A plausible explanation for single-level models that ignore FPCs, is the target population is so large (e.g., 4th grade students in the United States) and the sampling fraction is so small (i.e., less than 5%), that the need for a finite population correction is minimal. This explanation is less plausible for multilevel regression models generalizing to higher levels because those higher levels are more likely to be finite, or at least fewer, compared to the level-1 units. For example, the number of schools at level-2 cannot exceed the number of 4th grade students at level-1. FPCs should be applied to higher units of analysis as well, which are more likely to be

finite than their single-level alternatives. However, until recently "applications of FPC to HLM have not been thoroughly discussed in the literature" (Lai et al., 2018, p. 94).

Lai et al. (2018) fill that void in the literature and propose a FPC adjustment accounting for the size of the analytical sample relative to its population for fixed-effect standard errors in 2-level hierarchical linear models. Results from their simulation study indicate bias in unadjusted fixed-effect standard errors increases with the intraclass correlation coefficient, number of clusters (i.e., level-2 units), average cluster size (i.e., average number of level-1 units), and sample-population ratio (i.e., sampling fraction). Their FPC adjustment produced unbiased standard errors whereas omitting their adjustment resulted in overestimated standard errors and confidence intervals that are too wide. Unfortunately, as with many methods, their FPC adjustment does not work well in small samples and Lai et al. (2018) suggest the "proposed adjustment only be applied when the number of clusters is at least 30 with 10 or more observations in each cluster" (p. 108).

The estimation of standard errors is problematic for small samples regardless whether or not FPCs are applied. However, bootstrapped standard errors may provide more accurate results for applied researchers when dealing with few clusters (i.e., level-2 units) (Mass & Hox, 2004; Mass & Hox, 2005; McNeish & Stapleton, 2016a; McNeish & Stapleton, 2016b; Snijders & Bosker, 2012). Lai et al. (2018) suggest bootstrapping procedures may be an appropriate strategy for dealing with a small number of clusters and encourage future researchers to implement a bootstrap procedure and "evaluate its performance against" their proposed FPC adjustment (Lai et al., 2018, p. 108).

Bootstrapping procedures are a type of resampling method (Chernick, 1999; Chernick & LaBudde, 2011). Numerous procedures exist, but the simplest is Efron's (1979) nonparametric bootstrap (Davison & Hinkley, 1997). Nonparametric bootstrap methods involve resampling the observed data with replacement. The logic driving nonparametric bootstrapping procedures can be extended to resample from hierarchical data structures (Goldstein, 2011; van der Leeden, Meijer, & Busing, 2007).

Although many research studies have examined this problem of analyzing hierarchical data with a small number of clusters (Mass & Hox, 2005; McNeish & Stapleton, 2016a; McNeish & Stapleton, 2016b; McNeish, 2017), to the author's knowledge, there is currently no research comparing the efficiency of finite population bootstrapping techniques to the FPC adjustment in two-level hierarchical linear models with few clusters. Furthermore, much of the existing research about modeling data with a small number of clusters has focused solely on continuous predictors. Standard errors of binary predictors exhibit bias when the prevalence of those predictors is "highly discrepant" or unbalanced meaning that most of values fall within a single category (McNeish & Stapleton, 2016b, p. 302). Standard errors for a relatively balanced binary variable (e.g., gonosome) function similarly in terms of bias as continuous predictors. However, standard errors associated with a discrepant or unbalanced binary predictor (e.g., English language learner designation) exhibit bias, especially when based on fewer than 60 clusters (McNeish & Stapleton, 2016b).

The current body of literature on the applications of FPC to HLM is sparse and leaves many questions for applied researchers considering applying FPCs to their hierarchical data, especially when dealing with a few clusters and binary predictors. How

do unadjusted standard errors compare to FPC adjusted standard errors with a small number of clusters? How do unadjusted standard errors compare to FPC adjusted standard errors with large cluster sizes? Does the PFC adjustment produce unbiased standard errors in a small number of large clusters? How do standard errors derived from finite population bootstrapping techniques compare to standard errors with FPC adjustment for continuous predictors? Does the FPC adjustment work for binary predictors? Does the FPC adjustment work for unbalanced binary predictors? How do standard errors derived from finite population bootstrapping techniques compare to standard errors with FPC adjustment for binary predictors? Further research is needed to resolve these questions.

Purpose

The purpose of the current study is threefold. First, the current study evaluates the performance of Lai et al.'s (2018) FPC adjustment in two-level hierarchical linear models for a few number of large clusters. Next, the current study compares the performance of finite population bootstrapped standard errors to Lai et al.'s (2018) FPC adjustment. Finally, the current study examines the efficiency of Lai et al.'s (2018) FPC adjustment for standard errors associated with a binary predictor. Monte Carlo simulation methods were used to examine relative bias, mean square error (MSE), and root mean square error (RMSE) of the unadjusted and FPC adjusted fixed-effect standard errors. Coverage was an additional evaluation criterion for the binary predictor only.

The following design factors were manipulated: (a) number of clusters in the sample; (b) cluster size; (c) analysis method; (d) binary predictor ratio; and (e) binary predictor effect. The two factors relating to the binary predictor were isolated in their

own study because of anticipated convergence issues for models using binary predictors. Tables 1.1 and 1.2 summarize the design factors for the continuous predictors study and for the binary predictor study respectively.

Table 1.1. *Summary of Design Factors for the Continuous Predictors Study*

Factor	**Levels**
Number of clusters in the sample	1) $J = 20$ 2) $J = 30$ 3) $J = 60$
Cluster size	1) $n_j = 30$ 2) $n_j = 90$ 3) $n_j = 150$
Analysis method	1) No bootstrap; unadjusted 2) No bootstrap; FPC adjusted 3) Finite population bootstrap

Table 1.2. *Summary of Design Factors for the Binary Predictor Study*

Factor	**Levels**
Number of clusters in the sample	1) $J = 20$ 2) $J = 30$ 3) $J = 60$
Cluster size	1) $n_j = 30$
Analysis method	1) No bootstrap; unadjusted 2) No bootstrap; FPC adjusted 3) Finite population bootstrap
Binary predictor ratio	1) 50:50 2) 20:80
Binary predictor effect	1) $\gamma_{02} = .45$ 2) $\gamma_{02} = .20$

Study conditions were chosen based on their importance in past research relating to sufficient sample sizes for multilevel modeling, their use in other multilevel simulation studies, and their prevalence in applied educational settings. All levels of factors were fully crossed within each study, giving rise to a total of 63 simulation conditions (i.e., 27 conditions for the continuous predictors study and 36 additional conditions for the binary predictor study).

Significance

The FPC adjusted standard error estimates exhibited acceptable levels of relative bias across most conditions. However, FPC adjusted standard error estimates underestimated the empirical standard errors and were more biased than unadjusted standard error estimates when the number of clusters was 20. Relative bias in FPC adjusted standard error estimates was less than the relative bias in unadjusted standard error estimates when the number of clusters was 30. Consequently, it is suggested that the finite population adjustment be applied when then number of clusters is at least 30.

Standard error estimates for a balanced binary level-2 predictor functioned similarly in terms of bias as continuous predictors. The relative bias in FPC adjusted standard error estimates for a balanced predictor was smaller than the relative bias in unadjusted standard error estimates when the number of clusters was at least 30. More clusters are needed when estimating standard errors for an unbalanced binary predictor and it is recommended that the FPC adjustment be applied to unbalanced binary predictors' effects when the number of clusters is at least 60.

The current study has important implications for applied researchers when deciding whether to include FPCs in their hierarchical linear models. Ideally, applied researchers should consider their population of interest in the earliest stages of designing their studies and apply FPCs when appropriate because results of the current study demonstrate how unadjusted standard error estimates in HLMs are positively biased when a finite target population is ignored.

CHAPTER II. LITERATURE REVIEW

Applied researchers are interested in answering questions about a particular population. Unfortunately, a complete census of a population is not feasible in many situations because of time restrictions and budget constraints. As a result, researchers usually rely upon samples to draw conclusions about their population of interest. Conclusions may be biased if researchers fail to incorporate important sampling design features into their analyses (Kish, 1965; Little, 2004; Smith, 1994).

This chapter discusses topics essential to understanding how to incorporate a specific design feature (i.e., the sampling fraction or sample-population ratio) by using a finite population correction for fixed-effect standard errors in two-level hierarchical linear models as described in Lai et al. (2018). The discussion begins with an overview of sampling, specifically simple random sampling. The discussion continues with a summary of two competing theoretical frameworks for statistical inference (design- vs. model-based) and advantages of using an integrated or hybrid of the two framework. Next, an introduction to hierarchical linear modeling and an exploration of how it fits within the integrated framework is supplied. Examples of empirical studies in which the level-2 population is finite are provided, with an emphasis placed on studies conducted within educational settings. This chapter continues with the basic theory underlying finite population corrections before discussing current limitations of a FPC adjustment method for fixed-effect standard errors in 2-level HLMs (Lai et al., 2018). Finally, a multilevel bootstrapping technique is introduced as a plausible alternative to the FPC adjustment for modeling hierarchical datasets with few number of large clusters. Theoretical and empirical considerations are provided throughout this chapter.

Sampling

The goal of statistical inference is to make inferences about a population based a sample of its observations, but inferences may not be valid if researchers fail to account for features of their sampling design (Kish, 1965; Little, 2004; Smith, 1994). Historically, two distinct frameworks (design- vs. model-based) have been proposed in order to reach valid conclusions about a population based on a sample of observations. Recent research has focused on how those competing frameworks may be reconciled into a single, integrated framework (Lehmann, 1993; Little, 2004; Smith, 1994). However, it is necessary to define common sampling designs before discussing the advantages of the integrated framework.

Probability & Simple Random Sampling

The sampling design is the procedure by which the sample of units is selected from the population of interest (Thompson, 2012). Probability sampling refers to designs in which every element in the population has a known, nonzero probability of being included in the sample (Hansen, Madow, & Tepping, 1983; Kish, 1965; Pfeffermann, 1996). Probability sampling procedures share the following properties:

1. We are able to define the set of distinct samples, $S_1, S_2, \ldots, S_v$, which the procedure is capable of selecting if applied to a specific population.
2. Each possible sample S_i has assigned to it a known probability of selection π_i.
3. We select one of the S_i by a random process in which each S_i receives its appropriate probability π_i of being selected.
4. The method for computing the estimate from the sample must be stated and must lead to a unique estimate for any specific sample. We may declare, for example, that the estimate is to be the average of the measurements on the individual units in the sample. (Cochran, 1977, p. 9).

Probability sampling refers to designs in which every element in the population has a *known probability* of being included in the sample (Hansen et al., 1983; Kish, 1965;

Pfeffermann, 1996). Simple random sampling (SRS) is a specific type of probability sampling design in which *n* distinct units or elements are selected from the *N* units in the population such that each possible combination of *n* units is equally likely to be selected (Kish 1965; Thompson, 2012). The probability that the *i*th element is included in the sample (π_i) is the same for each element with SRS such that $\pi_i = n/N$ (Thompson, 2012). SRS takes elements from the population at random (Neyman, 1934) and is the most basic probability sampling method (Kish, 1965). All other procedures may be considered modifications (Kish, 1965). Under ideal conditions, SRS possesses the following properties: (a) equal probability of selection method; (b) unstratified selection; (c) random selection; (d) one-phase sampling; and (e) element sampling. Table 2.1 lists the idealized properties of probability sampling required for SRS and their alternatives as discussed in Kish (1965).

Table 2.1. *Properties of Probability Sampling and Alternatives to SRS*

	Idealized (SRS)	**Alternatives**
a)	Equal probability of selection for all elements	Unequal probabilities for different elements
b)	Unstratified selection	Stratified selection
c)	Random selection	Systematic selection
d)	One-phase sampling	Two-phase (double) & multiphase sampling
e)	Element sampling	Cluster sampling

Equal probability of selection method (Epsem) describes sampling designs in which all the population elements have equal inclusion probabilities. Loaded designs refer sampling designs with unequal probabilities for different elements (Kish, 1965). Loaded designs are not considered to be SRS because they are not Epsem (i.e., the inclusion probabilities are not the same for each unit). The idealized property of SRS, Espem, is desirable because it leads to self-weighing samples where the sample mean and

variance are unbiased estimators of the population mean and variance (Kish, 1965; Thompson, 2012).

SRS uses unstratified selection and is not possible when the population of interest is stratified or divided into subpopulations (Kish, 1965). Elements are selected from subpopulations, or strata, in stratified sampling designs. For example, suppose student learning outcomes differ in private and public institutions. Applied researchers interested in student learning outcomes may wish to sample from both public and private schools (i.e., from both strata). Stratified sampling is an alternative to SRS in which elements are selected from each strata. Stratified samples may or may not be selected at random. Stratified random sampling refers to designs in which elements are randomly selected from each strata (Shadish, Cook, & Campbell, 2002).

Random selection entails selecting elements or units by chance from the entire population (Shadish et al., 2002). Systematic selection is the alternative to random selection and denotes sampling of units in a sequence separated by an interval. Systematic selection involves selecting every *k*th unit, rather than selecting at random (Kish, 1965). Designs using systematic selection are not considered SRS because units are not selected at random.

The sample is selected directly from the population when using one-phase sampling. Two-phase and multiphase sampling methods are the alternatives (Kish, 1965). Two-phase, or double sampling refers to designs in which an initial sample of units is collected, and then a second sample is selected as a subset of the first (Thompson, 2012). Multiphase sampling methods refer to designs with more than two phases of selection (Kish, 1965).

Element sampling is the final idealized property of probability sampling required for SRS. The elements or individual units are the only sampling units in element sampling. Element sampling's alternative, cluster sampling, entails selecting groups of elements as sampling units (Kish, 1965). The sampling units are single elements when using element sampling whereas the primary sampling units (PSUs) are clusters of elements when using cluster sampling. Applied researchers often use multistage designs in which clusters are sampled in the first stage and elementary units are sampled in the final stage. These designs result in hierarchical data structures with the single elementary units in level-1 and the PSUs in the highest level-L. These are known as cluster sampling designs (Rabe-Hesketh & Skrondal, 2006). Using HLM to model hierarchical data structures resulting from cluster sampling is the focus of the current study and is expanded on in a later section.

Simple random samples may be collected with or without replacement. For sampling with replacement, selected n elements with equal probabilities of inclusion are returned to the sample and can be selected again. For sampling without replacement, selected elements are not returned to the sample (i.e., they can only be selected once). The remainder of this manuscript utilizes terminology consistent with Cochran (1977), Kish (1965), and Thompson (2012). Effectively, this manuscript uses SRS when referring to simple random sampling without replacement and unrestricted sampling when referring to simple random sampling with replacement.

The preceding discussion summarizes probability sampling, specifically, SRS. The ideal, most basic probability sampling design (SRS) is not always feasible and many datasets used in the social sciences are collected using schemes other than SRS (Lai et al.,

2018). The alternatives to SRS are listed in Table 2.1. Note that these alternatives may be combined. For example, it is possible to use systematic cluster sampling in which every *k*th cluster is included in the sample (Thompson, 2012). Also, clusters may be sampled with equal probabilities of inclusion (Epsem) or not. Finally, researchers may wish to sample from different strata within a cluster. Applied researchers need to carefully define their populations of interest and decide whether probability sampling is feasible because appropriate modes of statistical inference depend on the properties of their sampling designs.

Modes of Statistical Inference

Sampling is a more cost effective alternative to census (i.e., surveying an entire population), but the interest still lies in making inferences about an entire population and inferences may not be valid if researchers fail to account for features of their sampling design (Kish, 1965; Little, 2004; Smith, 1994). Therefore, applied researchers must consider the different types of samples (only random versus nonrandom or random) and apply the appropriate inferential framework (descriptive versus analytic) to different kinds of populations (finite versus infinite) in order to draw valid conclusions based on a sample (Sterba, 2009). (Random samples refers to samples with known probabilities of inclusion and nonrandom samples refers to samples with unknown probabilities of inclusion (Sterba, 2009).)

Competing Frameworks

Historically, researchers were forced to choose between two competing frameworks of statistical inference: design-based and model-based inference (Sterba, 2009). These frameworks offer different philosophies for reaching valid inferences about

a given population based on observations in the sample. Traditionally, these frameworks have been at "war" with one another (Kish, 1995/2003). More recent research suggests how these frameworks may be reconciled into a single integrated or hybrid framework (Lehmann, 1993; Little, 2004; Smith, 1994). However, understanding the advantages associated with utilizing an integrated framework warrants a preliminary discussion contrasting pure design- and pure model-based modes of inference.

Design-based. Pure design-based inference is also known as randomization, procedural, descriptive (Pfeffermann, 1993; Smith, 1994), representative (Neyman, 1934), deductive (Lehmann, 1993), fixed-population (Thompson, 2012), and probability sampling inference (Hansen et al., 1983; Smith, 1994). Design-based inference is based on the philosophy of Jerzy Neyman and focuses on sampling and features of the sampling design (Kish, 1995/2003). The goal of design-based framework is to make inferences about finite population parameters (Lai et al., 2018). For example, a sample mean $\bar{y}$ is an estimate of the finite population mean μ if using SRS under the design-based framework. The design-based framework is implemented using the following steps:

1. Specify frame, design, and scheme.
2. Estimate the finite population parameter of interest and its variance using known probabilities of inclusion (π_is).

Step 1 of the design-based framework requires specifying a random sampling frame, design, and scheme (Neyman, 1934). A sampling frame is a complete list of the PSUs in the population (Cochran, 1977). The sampling design is a procedure by which a sample of units is selected from the population and assigns known probabilities of selection to each sample that could be drawn from the frame (Thompson, 2012) and a sampling scheme is a mechanism for implementing the design (Sterba, 2009). Step 1 of design-

based inference possesses an inherent assumption (i.e., a finite, well-defined population of interest). The population must be *finite* in order to know the sampling frame because compiling a complete list of PSUs is not possible for populations of *infinite* size. Step 1 of design-based inference necessitates a well-defined sampling frame to make valid, useful inferences about a finite target population.

Step 2 of the design-based framework requires estimating a finite population parameter and its variance using the known probabilities of selection (π_is) for sampled units. Probability sampling is a fundamental requirement for design-based inference because of step 2 (Hansen et al., 1983). SRS (i.e., Epsem probability sampling) is preferred because it leads to self-weighting samples (Kish, 1965). The π_is are constant for SRS, and therefore, they may be ignored (Snijders & Bosker, 2012). However, Epsem probability sampling (SRS) designs are not always feasible and many samples are collected with positive, but unequal probabilities of selection (Lai et al., 2018).

The design-based framework must apply sampling weights in order to draw valid inferences from samples collected with unequal π_is (Kish, 1965; Snijders & Bosker, 2012). Sampling weights (w_is) reflect the inverse of the probability that any particular population unit is included in the sample. For single-level designs, the weight is $w_i = 1/\pi_i$ (Snijders & Bosker, 2012). Unequal selection probabilities are not explored in the current study and the simulation assumes SRS throughout all conditions.

The design-based framework cannot accommodate nonsampling errors (Sterba, 2009). Sampling errors are any errors caused by observing a sample instead of the entire target population (Särndal, Swensson, & Wretman, 1992). Nonsampling errors are all other errors and include errors due to nonobservation (i.e., nonresponse) and errors in

observations. Nonsampling errors occur when the frame does not correspond to the target population, when some of the PSUs cannot be observed, when variables are observed with measurement error, or when the actual probabilities of selection differ from those of the presumed design (Thompson, 2012). Assuming the population is divided into responders and nonresponders, then nonobservation is a form of selection bias (Thompson, 2012). Nonobservation or missing data cannot be accommodated by the design-based framework because the probabilities of selection required for step 2 cannot be known for missing observations. Errors in observations include measurement and processing errors (Särndal et al., 1992) and are another source of nonsampling error that cannot be accommodated by the design-based framework. Instead, accommodating missing data and measurement errors requires a model-based framework (Sterba, 2009).

Model-based. Pure model-based inference is also known as predictive (Smith, 1994), analytic (Pfeffermann, 1993), classical statistical (Pfeffermann, 1996), inductive (Fisher, 1955; Lehmann, 1993), stochastic-population (Thompson, 2012), and model-dependent inference (Hansen et al., 1983). Model-based inference is based on the philosophy of Sir Ronald A. Fisher and focuses on statistical theory and analysis (Kish, 1995/2003). The goal of the model-based framework is to make inferences about the parameters of a model generating the observed data in a sample (Lai et al., 2018). For example, a simple model for estimating a population mean would assume that each observation y_i is generated from a normal distribution with a mean μ and a variance σ^2. The model-based framework is implemented using the following steps:

1. Specify statistical model.
2. Impose parametric distributional assumption.
3. Meet conditionality principle.

Step 1 of the model-based inference requires formulating a statistical model to describe how the observed outcomes are thought to be generated (Fisher, 1922). The goal of a statistical model is to provide a link between the observed sample units and the unobserved units in the population (Sterba, 2009). An example of statistical model is a single-level regression model in which the dependent variable (y_i) is a function of an independent variable (x_i) and error (ε_i):

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (2.1)$$

(Sterba, 2009). Model-based inference treats the available observations as a sample from a hypothetical and infinite population (Fisher, 1922). The regression coefficients are the model parameters characterizing the infinite population. The hypothetical, infinite population is often referred to as the superpopulation (Cochran, 1977; Gelman & Hill, 2006; Hansen et al., 1993; Lai et al., 2018; Little, 2004). The remainder of this manuscript uses the terms *infinite population* and *superpopulation* interchangeably in order to remain consistent with the terminology posited by Lai et al. (2018).

Step 2 of the model-based inference requires imposing a parametric distributional assumption in order to treat the observations as realizations of a random variable (Fisher, 1922). The *iid* assumption is an example of a parametric distributional assumption often imposed in the model-based framework (Sterba, 2009). (The abbreviation *iid* refers to "identically and independently distributed random variables" (Kish, 1995/2003).) The *iid* assumption assumes that errors in the model are independent from each other and identically distributed with a mean of 0 and a variance σ^2. That is, $\varepsilon_i \sim iid\ N\ (0, \sigma^2)$ (Sterba, 2009). The model-based framework does not require SRS because random variation in the dependent variable is introduced by model assumptions (Johnstone,

1987). Although SRS is not a formal requirement of the model-based framework, by invoking parametric distributional assumptions in step 2, researchers are claiming the distribution of observed outcomes *does not differ* meaningfully from the distribution that would have been generated under SRS (Sterba, 2009). However, in many situations, a researcher's sample *does differ* meaningfully from what would have been generated under SRS. Dealing with these situations requires an additional step.

Step 3 of the model-based inference is referred to as Fisher's conditionality principle (Sterba, 2009). The conditionality principle requires conditioning models on any indicators or circumstances that may cause the sample distribution to meaningfully differ from the empirical distribution under SRS. According to the conditionality principle, nothing should distinguish a set of observations from any other set that could have been generated under the hypothetical model (Fisher, 1955; 1956). Meeting this principle ensures the infinite population is "subjectively homogeneous and without recognizable stratification" (Fisher, 1956, p. 33). In other words, meeting the conditionality principle ensures designs are uninformative, meaning selection probabilities do not contain any more information than the data upon which the design was based (Smith, 1994).

Fisher realized the sample distribution does meaningfully differ from the empirical distribution under SRS in certain circumstances (Sterba, 2009). These circumstances are referred to as informative designs, or designs in which there is an association between the residuals (ε_i) and the sampling design (Snijders & Bosker, 2012). Circumstances leading to informative designs occur when sampling units in the population are stratified or divided into categories, when sampling units are

disproportionately selected, and when sampling units are clustered or aggregated into groups (Sterba, 2009). The first two circumstances, stratification and disproportionate selection, are not explored in the current study and the simulation assumes no stratification and equal probabilities of selection throughout all conditions.

In the case of clustering, conditioning would amount to expanding a single-level regression model to allow coefficients to vary across clusters by including group indicator variables:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \varepsilon_{ij}, \tag{2.2}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}, \qquad \text{where } \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim iid\ N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{bmatrix}\right)$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

(Raudenbush & Bryk, 2002; Sterba, 2009). Informative designs due to clustering is central to the current study and is expanded on in the following sections when discussing hierarchical linear modeling. Note the hierarchical linear model presented in the equation above does not fall under the pure model-based framework. Rather, it is considered a hybrid framework because it integrates the clustering structure into a statistical model (Lai et al., 2018). The hybrid or integrated framework is discussed in later sections of this manuscript.

Design- vs. model-based inference. The preceding discussion summarizes the steps required for the design- and model-based frameworks of Neyman and Fisher respectively. These frameworks offer two distinct philosophical approaches for drawing inferences about a population given a sample of observations. Both are appropriate approaches provided applied researchers apply them correctly in a given setting and are aware that different types of samples (only random versus nonrandom or random) require

different frameworks (descriptive versus analytic) to reach valid inferences about different types of populations (finite versus infinite) (Sterba, 2009).

Both frameworks need to consider the observations (y_is) as realizations of a random variable in order to explain their variability (Sterba, 2009). In the design-based framework, y_is are considered as fixed unknown constants and randomness is deliberately imposed by design (Thompson, 2012). Consider a sample mean ($\bar{y}$) which varies from sample to sample. One sample's mean is higher than that of the population, another sample's mean is lower, and estimation of the population's mean improves with repeated sampling under the design-based framework. The probability distribution is determined by the sampling scheme and averages over all possible samples under the design-based framework (Smith, 1994). Probability sampling is a requirement for appropriate design-based inference because it is impossible to determine the probability distribution if the probabilities of selection (π_is) for sampled units are unknown (Hansen et al., 1983; Smith, 1994).

Design-based inference is only applicable to random samples because randomness under the design-based framework is an *empirical* state, dependent upon the sampling mechanism (Johnstone, 1987; Sterba, 2009). In contrast, randomness under the model-based framework is an *epistemic* state and is introduced through the imposition of parametric distributional assumptions (Johnstone, 1987; Sterba, 2009). The observations (y_is) are considered random variables under the model-based framework and applied researchers are claiming the distribution of observed values does not differ meaningfully from the distribution that would have been generated with empirical probability sampling by invoking distributional assumptions such as *iid* (Sterba, 2009). The model-based

framework may be applied to nonrandom (i.e., samples with unknown π_is) or random (i.e., samples with known π_is) samples as long as the parametric distributional assumptions hold and the conditionality principle is met.

Probability sampling (i.e., a scheme with known probabilities of inclusion) is required for design-based inference (Hansen et al., 1983; Smith, 1994). Although probability sampling is not a requirement for model-based inference, it is preferred because using probability sampling is the best strategy to ensure the design is uninformative (Smith, 1994). The model-based framework does require that the design is uninformative, meaning the sampling mechanism and selection probabilities are not related to the residuals. If an association exists between residuals and the sampling mechanism, then the design is informative, and pure model-based inferences risk being biased (Snijders & Bosker, 2012). Clustering is an example of one of the circumstances leading to informative designs (Sterba, 2009). Clustering would likely result in a violation of the *iid* assumption and cannot be accommodated by the pure model-based framework without conditioning on cluster indicator variables.

To sum up the two competing philosophies, the pure design-based framework requires probability sampling (i.e., known probabilities of inclusion) and allows for descriptive inference of finite populations based on only random samples (i.e., samples with known π_is). The pure model-based framework requires imposing distributional assumptions (e.g., *iid*) and allows for analytic inference of infinite populations based on samples with either known or unknown π_is (Sterba, 2009).

Proponents of the design-based approach, also known as randomizers, account for the survey design to provide inferences in large samples while avoiding the parametric

distributional assumption required for statistical modeling (Smith, 1994). One of the strengths of the design-based approach lies in its recognition of the finite population as a real entity (Smith, 1994). Descriptive finite population quantities such as means, proportions, and correlation coefficients are the target parameters for design-based inference (Pfeffermann, 1996). The design-based approach is common in epidemiology, sociology, health sciences, government census, and polling where the target parameters describe finite populations (Sterba, 2009).

Descriptive inference about finite populations is the goal of the design-based approach (Little, 2004). However, psychologists tend to be less interested in a particular finite population and more interested in constructing causal models to predict future behavior. As a result, psychologists tend to be proponents of the model-based framework (Sterba, 2009).

Proponents of the model-based approach, also known as modelers (Sterba, 2009), make inferences about model parameters such as expected values, variances, and regression coefficients (Pfeffermann, 1993). The parameters of the data generating mechanism (i.e., the superpopulation model) are the targets for model-based inference (Rabe-Hesketh & Skrondal, 2006). The strength of the model-based framework lies in its ability to draw inferences from nonrandom samples (i.e., samples without known π_is) (Sterba, 2009). The model-based framework does not require probability sampling because the probabilities of selection are irrelevant as long as the parametric distributional assumptions imposed hold true (Hansen et al., 1983). Stated another way, the model-based framework does not require probability sampling, but does require that the sampling mechanism is uninformative (Rubin, 1976; Sugden & Smith, 1984).

Nonrandom samples can result in biased model estimates if researchers fail to account for informative sampling mechanisms (Little, 2004).

Modelers criticize the pure design-based framework because it is limited to simple descriptive statistics and does not allow for causal inference (Sterba, 2009). A pure design-based approach cannot handle nonsampling errors. Rather, dealing with nonsampling errors requires using a model-based approach. Furthermore, the design-based approach is asymptotic (Little, 2004). As a result, the design-based approach is usually limited to large survey data (Lai et al., 2018) and may be unreliable in small samples (Snijders & Bosker, 2012). Finally, pure design-based inference is limited to populations with similar structures to that under study (Pfeffermann, 1993).

A pure-model based approach requires the imposition of strict parametric distributional assumptions (Sterba, 2009), which may not be realistic in many real-world settings. Fitting a model that actually approximates the population's values may not be practical because of heterogeneous populations encountered and complex sampling designs used in practice (Pfeffermann, 1993). Failure to account for sampling design features, such as clustering, likely violates the *iid* assumption. This violation may cause any models based on the sample to be very different from those in the population, leading to biased inferences (Pfeffermann, 1993). However, it may be impossible to include all the relevant design information and few researchers wish to do so because conditioning on all possible design variables and including their interactions complicates model specification (Pfeffermann, 1996).

A model is more or less efficient to the degree that the model is a good description of the real-world (Hansen et al., 1983). If a model is plausible representation

of the real-world, and sampling mechanism is independent of residuals, then design is irrelevant (Snijders & Bosker, 2012). However, a pure model-based approach may lead to biased inferences if the model is misspecified, or is not a plausible representation of reality (Little, 2004). Model misspecification is a pertinent issue for applied researchers because no model will include all the relevant variables (Pfeffermann, 1993). Moreover, a model is never more than an approximation and applied researchers cannot ever know if their models are "true" representations of the real-world (Snijders & Bosker, 2012, p. 219).

A major criticism of the model-based approach is that it has become dislodged from Fisher's philosophy because of complications in meeting the conditionality principle (Sterba, 2009). Recall the conditionality principle requires conditioning models on any design indicators or circumstances that may cause the sample distribution to meaningfully differ from the empirical distribution. Randomizers claim modelers cannot ever condition on all possible indicators, and therefore, cannot ever know if Fisher's conditionality principle is met (Pfeffermann, 1996). Applied researchers should condition on relevant design features whenever possible because models failing the conditionality principle risk misspecification and may result in biased inferences.

Both frameworks have limitations. The pure design-based framework is limited to descriptive inference for finite population parameters from samples with known probabilities of inclusion and is incongruent with causal inference (Sterba, 2009). The pure model-based framework, on the other hand, allows for causal or analytic inference for infinite population parameters with random or nonrandom samples, but is susceptible to bias from incomplete conditioning (Sterba, 2009). The model-based approach may be

used for nonrandom samples as long as the design is uninformative (Little, 2004). Neither pure approach is ideal given the criticisms noted earlier in this section. Because of those criticisms, applied researchers may benefit from implementing and integrated or hybrid approach capitalizing on the strengths of both competing frameworks. The FPC adjustment for two-level HLMs discussed later in this chapter is considered an integrated approach (Lai et al., 2018).

Integrated Framework

Traditionally speaking, the pure design- and model-based frameworks have been at "war" with one another (Kish, 1995/2003). However, "it is frequently the synthesis of existing ideas that can lead to great advances" (Smith, 2004, p. 6) and recent work has focused on how the two competing frameworks can be "reconciled" into an integrated or hybrid framework (Smith, 1994).

An integrated framework possess several advantages over using either a pure design- or model-based framework: (a) it can produce analytic statistics from complex samples without the need to condition on all of the sampling features; (b) it allows for analytic or descriptive inference about finite or infinite populations; (c) it can account for measurement error; and (d) it allows applied researchers to condition on some sampling features while ignoring others during model specification (Sterba, 2009).

The first advantage of using an integrated framework is its ability to account for the sampling design during model estimation. For example, assume the sampling design involves unequal probabilities of inclusion and clustering. The integrated framework adjusts for disproportionate selection and clustering during estimation rather than

conditioning on all selection variables when specifying the model and allows for analytic or causal inference (Sterba, 2009).

The next advantage of using an integrated framework is its ability to make analytic or descriptive inferences about finite or infinite populations. Traditionally, the pure design-based framework is limited to *descriptive* inference to *finite* populations and the pure model-based allows for *analytic* inference to *infinite* populations. An integrated approach permits both *analytic* and *descriptive* inference to either *finite* or *infinite* populations (Sterba, 2009).

Recall the pure design-based framework cannot accommodate nonsampling errors such as measurement error. Another advantage of using an integrated framework is its ability to handle measurement error (Sterba, 2009). Structural equation models use multiple observations to account for measurement error and serve as examples of a hybrid framework because they model nonsampling errors introduced via the sampling design. Measurement errors are not explored in the current study and the simulation assumes observations are measured without error throughout all conditions.

The final advantage of using an integrated approach is that it allows for researchers to condition on some complex sampling features during specification while adjusting for others during estimation. Imagine scenarios with unequal probabilities of selection and clustering resulting in hierarchical data structures. Modeling cluster indicator variables (based on the model-based framework) while adjusting estimates to account for disproportionate selection (based on the design-based framework) in these scenarios would require a hybrid approach (Sterba, 2009).

The integrated framework is a hybrid in the sense that it affords inference to both kinds of populations (finite and infinite) and does not require completely correct specification. The integrated framework is not considered a hybrid in the sense that it allows both random and nonrandom samples and is only applicable to samples with known probabilities of inclusion (Sterba, 2009).

Hierarchical linear models. If a model is a plausible representation of the real-world, and the sampling mechanism is independent of residuals, then design is irrelevant (Snijders & Bosker, 2012). However, this is unlikely to be the case because "it is seldom convenient or efficient to select a simple random sample" (Kish, 1995/2003, p. 14). Rigorous sampling schemes (i.e., Epsem probability sampling) tend to be the exception rather than the rule in applied research (Kish, 1995/2003) and much of the data used in the social sciences are collected using sampling schemes other than random sampling (Lai et al., 2018). Conventional estimates are not consistent if the design is informative meaning the probabilities of inclusion are related to the responses after conditioning on relevant covariates (Rabe-Hesketh & Skrondal, 2006). Inferences drawn from samples risk being biased if applied researchers fail to account for informative sampling designs (Little, 2004). As noted earlier, clustering is one of the circumstances leading to informative designs (Sterba, 2009). Cluster sampling methods result in hierarchical or multilevel data in which observations are nested within levels. For example, in education, 4th grade students may be nested within schools. The assumption that observations are independent of each other is often violated for hierarchical data. Single-level regression models are not appropriate for these datasets because ignoring the nesting often leads to biased estimates (Raudenbush & Bryk, 2002; Snijders & Bosker,

2012). As a result, approaches accounting for nested observations, such as hierarchical linear modeling, are needed to correctly model clustered data.

Hierarchical linear modeling is sometimes referred to multilevel modeling, mixed- effects modeling, random-effects modeling, or random-coefficient regression modeling Raudenbush & Bryk (2002). The remainder of this manuscript will use the abbreviation HLM to refer to hierarchical linear modeling in order to remain consistent with the terminology utilized by Lai et al. (2018) and Raudenbush and Bryk (2002). The HLM presented in this section is considered a hybrid approach because it incorporates sampling design features into a statistical model (i.e., it is a model conditional on cluster indicator variables). A brief introduction to HLM is provided below (for greater information see Raudenbush & Bryk, 2002).

Imagine a scenario in which there are $i = 1, 2, \ldots, n_j$ students (i.e., level-1 units) nested within $j = 1, 2, \ldots, J$ schools (i.e., level-2 units). The simplest HLM is equivalent to a one-way ANOVA with random effects:

$$\textit{Level-1:} \qquad y_{ij} = \beta_{0j} + \varepsilon_{ij}, \tag{2.3}$$

in which the dependent variable (y_{ij}) is a function of an intercept (β_{0j}) and error (ε_{ij}). We assume ε_{ij} is normally distributed with a mean of 0 and a unknown variance σ^2 for every level-1 unit i within each level-2 unit j. β_{0j} is the mean of the dependent variable for the level-2 unit j (i.e., $\beta_{0j} = \mu_{y_i}$). A lack of independent observations (i.e., dependency) is explicitly modeled by allowing to β_{0j} vary across level-2 units with the following equation:

$$\textit{Level-2:} \qquad \beta_{0j} = \gamma_{00} + u_{0j}, \tag{2.4}$$

where γ_{00} represents the grand mean of the dependent variable in the population and u_{0j} represents the random effect associated with level-2 unit j. We assume u_{0j} is normally distributed with a mean of 0 and a unknown variance τ_{00}. Substituting Equation 2.4 into Equation 2.3 yields the following combined model:

$$\textit{Combined:} \quad y_{ij} = \gamma_{00} + u_{0j} + \varepsilon_{ij}, \tag{2.5}$$

in which variance of the dependent variable is $\text{Var}\left(y_{ij}\right) = \text{Var}\left(u_{0j} + \varepsilon_{ij}\right) = \tau_{00} + \sigma^2$. The model presented with Equation 2.5 is often referred to as a fully unconditional model because no predictors are specified at either level (Raudenbush & Bryk, 2002).

Estimating a fully unconditional model, or a one-way random effects ANOVA model presented in Equation 2.5, is an essential preliminary step in hierarchical data analyses because doing so provides information about the variability of the dependent variable at each level. The σ^2 parameter represents the within group (i.e., level-1) variability and τ_{00} represents the between-group (i.e., level-2) variability. The intraclass correlation coefficient (ICC) represents the proportion of variance in the dependent variable that is between level-2 units and is represented by the following equation:

$$\text{ICC} = \rho = \tau_{00}/(\tau_{00} + \sigma^2)\ . \tag{2.6}$$

The ICC measures the homogeneity of clusters (Thomas & Heck, 2001). Stated another way, the ICC serves as an estimate of the dependency in observations due level-2 cluster membership (i.e., the degree to which the *iid* assumption is violated). The ICC is a useful diagnostic for deciding whether HLM is necessary and should be zero when observations are independent from each other. An ICC of zero indicates no variation in the outcome of interest across level-2 units (i.e., no dependency) and traditional single-level techniques can be used to analyze the data. However, inferences drawn from single-level

techniques risk being biased if the ICC is not zero. If the ICC is positive, then a violation of the *iid* assumption has occurred and HLM must be used in order to draw valid inferences from the data (Peugh, 2010).

Other than serving as a diagnostic tool to estimate the ICC, the fully unconditional model is of little substantive interest to applied researchers because it does not afford any predictive inference (i.e., no independent variables are specified in the model to explain variation in the outcome). Applied research questions often seek to explain that variation and doing so requires specifying model predictors. Recall the example of a single-level regression model represented by Equation 2.1. In the presence of clustering, fulfilling Fisher's conditionality principle amounts to expanding Equation 2.1 to include cluster indicator variables (Sterba, 2009) and yields the following equation for a random-coefficients regression model:

$$\begin{aligned} &\textit{Level-1:} && y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \varepsilon_{ij} \\ &\textit{Level-2:} && \beta_{0j} = \gamma_{00} + u_{0j} \\ & && \beta_{1j} = \gamma_{10} + u_{1j} \\ &\textit{Combined:} && y_{ij} = \gamma_{00} + u_{0j} + \gamma_{10}x_{ij} + u_{1j}x_{ij} + \varepsilon_{ij}, \end{aligned} \tag{2.7}$$

where β_{0j} is the level-1 intercept, β_{1j} is the level-1 slope, $\gamma_{00}$is the average intercept across level-2 units, γ_{10} is the average slope across level-2 units, u_{0j} is the j^{th} level-2 unit's unique effect on the intercept, u_{1j} is the j^{th} level-2 unit's unique effect on the slope. The dispersion of the level-2 effects (i.e., u_{0j} and u_{1j}) can be represented as a variance-covariance matrix: $\text{Var}\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{bmatrix} = \mathbf{T}$, where the parameters τ_{00}, τ_{11}, and τ_{10} represent the unconditional variances in the level-1 intercepts, slopes, and the

unconditional covariance between level-1 intercepts and slopes respectively (Raudenbush & Bryk, 2002). The intercept (β_{0j}) and slope (β_{1j}) in Equation 6 are allowed to vary across level-2 units as indicated by the subscript j.

Equation 2.7 includes only one, level-1 predictor x_{ij}. Suppose the outcome variable y_{ij} is affected by a level-2 binary predictor w_j that is dummy coded (e.g., public = 0 vs. private =1 schools). Expanding Equation 2.7 to include a level-2 predictor yields the following:

$$\textit{Level-1:} \quad y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \varepsilon_{ij} \tag{2.8}$$

$$\textit{Level-2:} \quad \beta_{0j} = \gamma_{00} + \gamma_{01}w_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}w_j + u_{1j}$$

$$\textit{Combined:} \quad y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + \gamma_{01}w_j + \gamma_{11}x_{ij}w_j + u_{0j} + u_{1j}x_{ij} + \varepsilon_{ij}.$$

Because of the dummy coded level-2 predictor in in Equation 6, γ_{00} is the mean outcome (i.e., intercept) and $\gamma_{10}$is the mean slope for level-2 units coded 0. γ_{01} is the mean difference in intercepts and γ_{11} is the mean difference in slopes between level-2 units coded 0 and 1. u_{0j} and u_{1j} are the j^{th} level-2 unit's unique effect on the mean outcome and slope, respectively, conditioning on w_j.

Equation 2.8 contains only a single level-1 predictor and a single level-2 predictor. However, the HLM framework can be expanded to include any number of *Q* level-1 predictors and *S* level-2 predictors with the following equation:

$$\textit{Level-1:} \quad y_{ij} = \beta_{0j} + \sum_{q=1}^{Q}\beta_{qj}\,x_{qij} + \varepsilon_{ij} \tag{2.9}$$

$$\textit{Level-2:} \quad \beta_{qj} = \gamma_{q0} + \sum_{s=1}^{S_q}\gamma_{qs}\,w_{sj} + u_{qj}.$$

Equation 2.9 represents a general form for two-level HLMs. The general form and the other HLMs discussed in this section assume the following:

1. $\varepsilon_i \sim iid\ N\ (0, \sigma^2)$.
2. $u_{qj} = \left(u_{0j}, \ldots, u_{Qj}\right) \sim iid\ N\ (0, \mathbf{T})$.
3. $\mathrm{Cov}\left(x_{qij}, \varepsilon_{ij}\right) = 0$.
4. $\mathrm{Cov}\left(w_{sj}, u_{qj}\right) = 0$.
5. $\mathrm{Cov}\left(\varepsilon_{ij}, u_{qj}\right) = 0$.
6. $\mathrm{Cov}\left(x_{qij}, u_{q'j}\right) = 0$.
7. $\mathrm{Cov}\left(w_{sj}, \varepsilon_{ij}\right) = 0$. (Raudenbush & Bryk, 2002, p. 225).

The general form for a HLM presented in Equation 2.9 is an example of a hybrid approach because it incorporates sampling design features due to clustering into a statistical model (Lai et al., 2018). Note this manuscript focuses solely on two-level HLMs. However, the principles discussed in this section may be expanded to incorporate any *L* levels of nesting.

HLMs provide applied researchers with a strategy for modeling dependency among observations due to clustering. As long as those models are correctly specified, the predicted variance components across samples that could be generated by the model ($\begin{bmatrix} \widehat{\tau_{00}} & \\ \widehat{\tau_{10}} & \widehat{\tau_{11}} \end{bmatrix}$) can be used to make inferences about the target parameters ($\begin{bmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{bmatrix}$) for an *infinite* population (Sterba, 2009). However, generalizations from the HLMs discussed in this section do not apply to *finite* populations. Applied researchers wishing to generalize findings based on their HLMs to finite populations need to consider incorporating an additional design feature, the sampling fraction or sample-population ratio, into their statistical models by using a FPC adjustment. Situations requiring FPC adjustments are described in the following section.

Why use Finite Populations Corrections?

FPCs are seldom utilized in single-level studies. One plausible reason for the lack of FPCs in single-level studies is that the target populations in those studies tend to be so large that FPC is unnecessary. "The omission of FPC for single-level studies may be justified by referencing an extremely large finite population" (Lai et al., 2018, p. 96) because the sample-population ratio tends to be less than 5% in single-level studies. However, this is unlikely to be the case when generalizing findings to level-2 units (e.g., schools) because the number of level-2 units is fewer and more likely to be finite relative to the elements in level-1 (e.g., students). The popularity of the model-based approach, which assumes the target population is hypothetical and infinite, is another reason for the lack of FPCs in single-level studies (Lai et al., 2018). However, as noted earlier, inferences to a finite population cannot be drawn from a pure model-based approach. Ultimately, the decision to use FPCs depends the target of inference.

Target of Inference and Fixed vs. Random Effects

Applied researchers wanting to analyze observations in groups must decide to treat cluster or group effects as either fixed or random. This section defines fixed and random effects, compares their common uses, and offers some suggestions for deciding whether to treat a particular effect as fixed or random.

Generally speaking, fixed effects are constant across individuals, whereas random effects vary (Kreft & De Leeuw, 1998). The intercept (β_0) and slope (β_1) in a single-level regression model are fixed effects (Snijders & Bosker, 2012). The HLM framework presented earlier allows those coefficients to vary between groups, as indicated by the subscript j in Equations 2.7-2.9. The regression coefficients β_{0j} and β_{1j} in a HLM are

sometimes called random effects (Gelman & Hill, 2006). For this reason, HLMs are sometimes referred to as random-effects models or random-coefficient regression models (Raudenbush & Bryk, 2002).

The same effect may be treated as either fixed or random (Gelman & Hill, 2006; Searle, Casella, & McCulloch, 1992; Snijders & Bosker, 2012). Unfortunately, "clear answers to the question 'fixed or random?' are not necessarily the norm" (Searle et al., 1992, p. 15). Rather, the appropriate interpretation of effects (fixed versus random) depends on the focus of the statistical inference, the nature of the groups included in the sample, and the population (Snijders & Bosker, 2012).

First, applied researchers must consider the type of statistical inferences they want to draw form their data. If researchers are only interested in within-group differences, then fixed effects are appropriate. If researchers are interested in differences between groups, then random effects should be used (Snijders & Bosker, 2012).

Researchers must also account for the nature of the groups included in the sample when deciding to treat an effect as fixed or random. Effects should be fixed if the J groups are regarded as unique categories (e.g., gonosome) and the researcher want to draw conclusions pertaining to those specific categories (Snijders & Bosker, 2012). Effects should be fixed if groups are interesting in themselves, that is, attention is "fixed" upon the groups in the model and "no others" (Searle et al., 1992, p. 7). However, if groups are regarded as a random sample from population (e.g., a sample of schools) and the researcher wishes to generalize findings to all the groups in that population (i.e., the other schools not included in the sample), then random effects are appropriate (Snijders & Bosker, 2012). Practical issues such as the number of clusters J must also be

considered when making the choice of treating group effects as fixed or random (Lai et al., 2018) because parameter estimations with random group effects risk being biased when J is fewer than 30 (Maas & Hox, 2004).

Every practical statistician must ask "of what population is this a random sample?" (Fisher, 1922, p. 313). Answering this question informs the decision to treat a particular effect as fixed or random. Green and Tukey (1960) suggest treating group effects as fixed when sample exhausts the population (i.e., $f = 1$) and random when the sample is a small, negligible part of the population. A sample is a small, negligible part of the population when $f < .05$ (Cochran, 1977).

Green and Tukey's (1960) suggestion leaves "open the question of what to do with a large but not exhaustive sample" (Gelman & Hill, 2006, p. 245). Consider the following example discussed in Gelman and Hill (2006) in which a researcher has collected data from 20 of the 50 states. The grouping variable "states" may be treated as either fixed or random depending on the target of generalization. Treating state group effects as fixed suggests there is no underlying population distribution of interest and any inferences drawn are limited to only those 20 states included in the sample. In contrast, treating state group effects as random suggests generalizing to an infinite number of states not included in the sample (e.g., provinces in Canada or cantons in Switzerland) and likely overestimates the sampling error. However, suppose the researcher would like to generalize his or her findings beyond the 20 states in the sample to the entire United States, but not to other "states." Neither fixed nor random effects are ideal in this example. Instead, it may be more meaningful to generalize to the finite population of 50

states, rather than limit findings to only the 20 states sampled using fixed effects or assume an infinitely large superpopulation using random effects (Gelman & Hill, 2006).

To summarize, fixed effects are usually attributable to a finite set of groups, whereas random effects are assumed to be from an infinite population or one large enough to be considered infinite for most practical purposes, as is the case in most single-level studies (Searle et al., 1992). Applied researchers need to consider the characteristics of their population in order to justify treating a grouping variable a fixed or random (Lai et al., 2018). This fixed versus random distinction is problematic in situations where one has sampled a non-negligible portion of level-2 units from a finite population. The use of the two-level HLM with FPCs approach described later in this chapter is situated "between the fixed and random ends of the group effect continuum" and is most appropriate in situations where the population of interest is clearly defined with a limited size (Lai et al., 2018, p. 97). Examples of empirical studies using HLM where the target of generalization at level-2 can be considered a finite population are reviewed in the following section.

Multilevel Studies in which the Population is Finite

In recent decades, hierarchical linear modeling has received increasing interest in the social sciences because of the realization that empirical studies in these disciplines often relate to data with a hierarchical structure (Maas & Hox, 2004; Raykov, 2010; Raykov et al., 2016). However, the theory behind hierarchical linear modeling assumes observations are sampled from a population of infinite size and little attention is given to situations where observations are a subset of a limited, finite, target population (Lai et al., 2018). This assumption poses a problem for applied researchers analyses because

standard errors are overestimated when the sample size exceeds the population size by as little as 5%. FPCs should be applied to adjust those standard errors in situations in which the sampling fraction or sample-population ratio is not negligible or exceeds 5% (Cochran, 1977).

Applied researchers adopting a traditional HLM approach are implying that their populations are infinite hypothetical entities (Snijders & Bosker, 2012). However, sometimes observations analyzed by HLM come from populations that are finite in size. Finite populations play an obvious role in cross-cultural research treating countries as the level-2 units (Lai et al., 2018). For example, Mostafa (2013) examined intentions to protect the environment based on observations from 25 countries. Peretz and Fried (2012) explored performance appraisal and organizational absenteeism across 21 countries. Rockstuhl, Dulebohn, Ang, and Shore (2012) studied leader-member exchange based on 28,587 individuals nested within 23 countries. The use of traditional HLM techniques is not appropriate in these examples because the number of countries in the world is not infinite. According to worldatlas.com, the United Nations recognizes 195 sovereign countries, which is far fewer than the infinite hypothetical superpopulation assumed by traditional HLM techniques.

Finite populations are not solely limited to national or cross-cultural research. Other examples include a study of 165 companies out of the 269 listed on the Swiss Stock Exchange (Nielsen, 2009); a study of 4,459 subsidiaries representing 40% of the total number of subsidiaries in Japan (Mani, Antia, & Rindfleisch, 2007); and a study of players from 3,569 out of 5,260 track and field clubs in North-Rhine Westphalia, Germany (Swierzy, Wicker, & Breuer, 2018). Each of these studies used traditional

HLM techniques based a non-negligible number of level-2 units sampled from a well-defined finite population.

Finite populations also play a role in many educational settings in which students are nested within a finite set of level-2 units (e.g., schools). For example, Maxwell, Reynolds, Lee, Subasic, and Bromhead (2017) studied literacy and numeracy achievement of students from 17 out of 19 schools in a district in Australia. Montague, Krawec, Enders, and Dietz (2014) examined mathematical problem solving skills of students with learning disabilities sampled from 40 out of the 78 schools in Miami-Dade County. Oberle, Schonert-Reichl, and Zumbo (2011) explored life satisfaction of students in 25 schools in Western Canada. Juvonen, Wang, and Espinoza (2011) studied the effects of bullying on the academic performance of students in 11 middle schools located in Los Angeles. Thrash and Warner (2016) examined substance use in adolescents from 287 schools in Nebraska.

All of these examples used HLM. Some of them are more explicit when stating their target populations than others, but none of them used FPC. Closer examination of these examples reveals their target populations are finite. Failing to account for the finite populations in these studies may have resulted in in overestimated standard errors of the regression coefficients. The following section summarizes topics essential to understanding finite population corrections.

Finite Population Corrections

"Inferential disasters can be avoided by selecting models that are attentive to design features" (Little, 2004, p. 551). The sampling fraction or sample-population ratio is one of those design features deserving attention from applied researchers because

estimates are biased when the sample size exceeds the population by as little as 5% (Cochran, 1977). FPCs must be applied in these situations to ensure accurate inference (Lai et al., 2018). The following discussion serves as a brief introduction to FPCs and how to calculate variance for finite populations, summarizes a method to compute adjusted standard errors from two-level HLMs for finite populations as described in Lai et al. (2018), and notes the current limitations of their method.

Estimating Finite Population Parameters

Consider a sample of size *n* taken using SRS from a population of size *N*. For SRS, the population mean μ is the average of the observations in the entire population such that

$$\mu = \frac{1}{N} \sum_{i=1}^{N} y_i \,, \tag{2.10}$$

and the sample mean $\bar{y}$ is the average of the observations in the sample such that

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{2.11}$$

(Thompson, 2012). Recall that SRS is an Espem, which is desirable because it leads to self-weighing samples where the sample mean $\bar{y}$ and variance s^2 are unbiased estimators of the population mean μ and variance σ^2 (Kish, 1965; Thompson, 2012). The equations for calculating variance of a mean and its estimate for infinite populations differ from those for finite populations (Cochran, 1977; Thompson, 2012) and are presented in Table 2.2.

Table 2.2. *Infinite and Finite Population Equations*

Equation	**Infinite**	**Finite**
Population variance	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2$	$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2$

Sample variance	$s^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$
Variance of the mean	$\mathrm{Var}(\bar{y}) = \frac{\sigma^2}{n}$	$\mathrm{Var}(\bar{y}) = (\frac{N-n}{N})\frac{\sigma^2}{n}$
Unbiased estimate of variance of the mean	$\widehat{\mathrm{var}}(\bar{y}) = \frac{s^2}{n}$	$\widehat{\mathrm{var}}(\bar{y}) = (\frac{N-n}{N})\frac{s^2}{n}$

The variance of the mean for an infinite population is σ^2/n which can be estimated using s^2/n (Cochran, 1977; Thompson, 2012). For finite populations, estimates of the sampling variance depends on both the sample (*n*) and the population size (*N*) and, as seen in Table 2.2, the equations for variance of the mean and its estimate are modified for finite populations by including the correction factor ($\frac{N-n}{N}$). This correction factor is usually referred to as the finite population correction (FPC) factor where

$$\mathrm{FPC} = \frac{(N-n)}{N} = 1 - \frac{n}{N} = 1 - f \tag{2.12}$$

(Thompson, 2012). The FPC ranges from 0 to 1. The smaller the sampling fraction (*f* = *n*/*N)*, the closer FPC is to 1. Substituting $1 - f$ for $\frac{(N-n)}{N}$ to represent $\widehat{\mathrm{var}}(\bar{y})$ in terms of the sampling fraction yields

$$\widehat{\mathrm{var}}(\bar{y}) = (1 - f)\frac{s^2}{n}. \tag{2.13}$$

The square root of the variance of the estimator is its *SE* such that

$$SE_{\bar{y}} = \sqrt{\widehat{\mathrm{var}}(\bar{y})} = \sqrt{(1 - f)\frac{s^2}{n}} = \frac{s}{\sqrt{n}}\sqrt{1 - f} \tag{2.14}$$

(Cochran, 1977; Thompson, 2012). The size of the population has a negligible effect on the standard error of the sample mean as FPC approaches unity, as shown with Equation

2.14. Ignoring finite populations may results in biased standard errors of the estimates when *f* exceeds .05 in single-level studies (Cochran, 1977).

Two-stage sampling. Many datasets used in the social sciences are collected using some alternative to SRS, such as cluster sampling (Lai et al., 2018); see Table 2.1. Typical two-stage cluster sampling techniques involve randomly sampling clusters from a level-2 population. Sometimes the level-2 population is finite, as evident in the examples discussed above.

Consider a finite population size of J_{pop} clusters, each with N_j elements, from which a random sample of J clusters is selected, with n_j elements drawn from each cluster. If is the value y_{ij} is the value obtained for the i^{th} element in the j^{th} cluster, then

$$\overline{y_j} = \frac{\sum_{i=1}^{n_j} y_{ij}}{n_j} \tag{2.15}$$

is the sample mean in the j^{th} cluster and

$$\bar{\bar{y}} = \frac{\sum_{j=1}^{J} \overline{y_j}}{J} = \frac{\sum_{j=1}^{J} \sum_{i=1}^{n_j} y_{ij}}{J n_j} \tag{2.16}$$

is the overall sample mean (Cochran, 1977). Applied researchers must estimate the variance at each level and add those terms together to get an overall estimate of the variance from two-stage sampling, as seen in the following equation. If units are selected at random, then an unbiased estimate of $\mathrm{Var}\,(\bar{\bar{y}})$ under two-stage sampling with FPC applied is

$$\widehat{\mathrm{var}}(\bar{\bar{y}}) = \frac{1-f_1}{J} {s_1}^2 + \frac{f_1(1-f_2)}{n_j J} {s_2}^2, \tag{2.17}$$

where

$${s_1}^2 = \frac{\sum_{j=1}^{J} (\overline{y_j} - \bar{\bar{y}})^2}{J-1} \tag{2.18}$$

is the variance among cluster means,

$$s_2{}^2 = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij}-\overline{y_j})^2}{J(n_j-1)} \tag{2.19}$$

is the variance among individual elements within clusters,

$$f_1 = J/J_{pop} \tag{2.20}$$

is the sampling fraction in the first stage, and

$$f_2 = n_j/N_j \tag{2.21}$$

is the sampling fraction in the second stage (Cochran, 1977). Substituting Equations 2.18-2.21 for $s_1{}^2$, $s_2{}^2$, f_1, and f_2 into Equation 2.17 yields the following alternative form,

$$\widehat{\text{var}}(\bar{\bar{y}}) = \frac{J_{pop}-J}{J_{pop}}\frac{1}{J}\frac{\sum_{j=1}^{J}(\overline{y_j}-\bar{\bar{y}})^2}{J-1} + \frac{J}{J_{pop}}\frac{N_j-n_j}{N_j}\frac{1}{Jn_j}\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij}-\overline{y_j})^2}{J(n_j-1)}. \tag{2.22}$$

Applying FPC to the variance components at each level and expressing the standard errors of regression coefficients in terms of the finite population adjusted variance components is one way to account for finite populations (Lai et al., 2018). The following section summarizes a method used to compute adjusted standard errors from two-level HLMs for finite populations as described in Lai et al. (2018).

FPC for the General Two-Level Linear Mixed Model

The combined model represented by Equation 2.8 has the same form as the general two-level linear mixed model,

$$\mathbf{y} = \mathbf{X\gamma} + \mathbf{Zu} + \boldsymbol{\varepsilon}, \tag{2.23}$$

where is **y** a vector of outcomes, **X** is a design matrix with *N* rows and *p* + 1 columns (i.e., one column for the intercept and one additional column for each predictor *p*), **γ** is a vector for fixed effects parameters, **Z** is a design matrix with *N* rows and *q* + 1 columns for the *q* variables hypothesized to have random effects, **u** is a vector containing the *q* + 1

level-2 random effects (i.e., $\mathbf{u} = [u_0, \ldots, u_q]$), and $\boldsymbol{\varepsilon}$ is a vector containing the level-1 error terms (Dedrick et al., 2009; Lai et al., 2018). The general two-level linear mixed model presented in Equation 2.23 assumes Cov($\mathbf{u}$, $\boldsymbol{\varepsilon}$) = 0, E($\boldsymbol{\varepsilon}$) = E($\mathbf{u}$) = 0, Var($\boldsymbol{\varepsilon}$) = $\sigma^2\mathbf{I}$ where $\mathbf{I}$ is an $N \times N$ identity matrix, and Var($\mathbf{u}$) = $\mathbf{T}$. Recall that for models with a random intercept and one random slope

$$\text{Var}(\mathbf{u}) = \text{Var}\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} \tau_{00} \\ \tau_{10}\ \tau_{11} \end{bmatrix} = \mathbf{T}. \tag{2.24}$$

The population variance of $\mathbf{y}$ is

$$\text{Var}(\mathbf{y}) = \text{Var}(\mathbf{Zu}) + \text{Var}(\boldsymbol{\varepsilon}) = \mathbf{ZTZ'} + \sigma^2\mathbf{I} = \mathbf{V} \tag{2.25}$$

for models in the form of Equation 2.23. Note that $\mathbf{Z'}$ is the transpose of $\mathbf{Z}$. Estimates of fixed effects coefficients are contained in the vector

$$\hat{\boldsymbol{\gamma}} = (\mathbf{X'V^{-1}X})^{-1}\mathbf{X'V^{-1}y}\,, \tag{2.26}$$

where $\mathbf{V^{-1}}$ is the inverse of $\mathbf{V}$ with

$$\text{Var}(\hat{\boldsymbol{\gamma}}) = (\mathbf{X'V^{-1}X})^{-1} \tag{2.27}$$

(Lai et al., 2018; Snijders & Bosker, 1993). The standard errors (*SE*s) of the regression coefficients are the square roots of the diagonal elements of $\text{Var}(\hat{\boldsymbol{\gamma}})$. Equations 2.23-2.27 assume populations of *infinite* size and may not be appropriate in situations where the sampling fraction *f* exceeds 5%. Handling samples with larger sampling fractions taken from *finite* populations requires an adjustment to the *SE*s.

Equation 2.25 suggests that the population variance $\mathbf{V}$ can be explained by a combination of level-1 (i.e., $\sigma^2\mathbf{I}$) and level-2 components (i.e., $\mathbf{ZTZ'}$) (Lai et al., 2018). As demonstrated in Cochran (1977) and Thompson (2012), applying FPC to the variance components at each level and expressing the *SE*s of the regression coefficients in terms of

the adjusted variance components is one way to account for finite populations. With FPC applied, the population variance of $\mathbf{y}$ is

$$\boldsymbol{V}^{*} = \text{FPC}_2 \times \mathbf{ZTZ}' + \text{FPC}_1 \times \sigma^2\mathbf{I} = \mathbf{ZT}^{*}\mathbf{Z}' + \sigma^{2^{*}}\mathbf{I}, \tag{2.28}$$

where

$$\mathbf{T}^{*} = \text{FPC}_2 \times \mathbf{T},$$

$$\sigma^{2^{*}} = \text{FPC}_1 \times \sigma^2,$$

and FPC_2 and FPC_1 are the finite population correction factors for level-2 and level-1, respectively (Lai et al., 2018). Consequently, the variance of the fixed effects regression coefficients represented in Equation 2.27 with FPC applied is

$$\text{Var}^{\text{FP}}(\hat{\boldsymbol{\gamma}}) = (\mathbf{X}'\mathbf{V}^{*-\mathbf{1}}\mathbf{X})^{-1} \tag{2.29}$$

(Lai et al., 2018). The standard errors of regression coefficients adjusted for finite populations (SE^{FP}) are the square roots of the diagonal elements of $\text{Var}^{\text{FP}}(\hat{\boldsymbol{\gamma}})$.

Lai et al. (2018) demonstrates how applied researchers can compute *SE*s adjusted for finite populations to obtain correct inference for the fixed effects from their HLM analysis. Their FPC method is considered to be a hybrid approach because it incorporates an additional design feature (i.e., the sampling fraction or sample-population ratio) into a more traditional model-based framework with random group effects. Lai et al. (2018) concludes with a Monte Carlo simulation to compare the performance of their FPC adjusted standard errors (SE^{FP}s) to unadjusted standard errors (SE_0s) under conditions with unbalanced cluster sizes. A summary of their design factors is presented in Table 2.3.

Table 2.3. *Summary of Design Factors from Lai et al. (2018)*

Factor	**Levels**
Data generation	1) Random intercept only 2) Random intercept & slope
Sample-population ratio	1) $P = .05$ 2) $P = .10$ 3) $P = .25$ 4) $P = .50$
Number of clusters in the sample	1) $J = 20$ 2) $J = 30$ 3) $J = 50$ 4) $J = 100$
Average cluster size	1) $\overline{n_j} = 5$ 2) $\overline{n_j} = 10$ 3) $\overline{n_j} = 25$
ICC	1) $\rho = .05$ 2) $\rho = .20$ 3) $\rho = .35$

Results from Lai et al.'s (2018) simulation demonstrates how failing to account for finite populations may lead to biased inferences. The degree of bias increases as *P*, *J*, $\overline{n_j}$, or ρ increases. The sampling fraction or sample-population ratio (*P*) explains the most variability in bias in the fixed effects. Their FPC adjustment method removed much, but not all of the bias for the fixed effects. However, their FPC adjustment is still considered an improvement because the SE^{FP}s were closer to the empirical estimates than the SE_0s across most conditions. FPC does not affect standard errors for the level-1 fixed effects in the random intercept only model because those *SE*s are not functions of the level-2 variance components. In contrast, FPC does play a role for the random intercept and slope model because the *SE*s in that model are functions of the random slope variance in level-2 (Snijders, 2005). Note Lai et al. (2018) did examine the effects of grand-mean and group-mean centering. However, those results were aggregated

because the differences between grand-mean and group-mean centering were negligible across all conditions.

Results from their simulation suggest standard errors in a HLM without the FPC adjustment are biased when the assumption of an infinite population at level-2 is violated (Lai et al., 2018). Applying their proposed adjustment produced acceptable *SE*s across most of the simulated conditions.

Current limitations of FPC for HLMs. The FPC adjustment method described in Lai et al. (2018) is most appropriate in situations where populations are well defined and limited in size. However, inferences from their simulation are limited to only the specific factors listed in Table 2.3. As noted in the examples of multilevel studies discussed above, finite populations may be problematic for cross-cultural research, organizational research, and other areas of social science research including education. Some of the samples used in applied research, specifically in education, fall outside the simulated conditions in Lai et al. (2018).

Lai et al.'s (2018) FPC adjustment method "produced acceptable *SE*s" in their simulated conditions (p. 106), but their simulation only examined the effects of *continuous* predictors. Many social research questions may require dealing with *binary* predictors. Gonosome (heterogametic vs. homogametic) and English language learner designation (no vs. yes) are examples of level-1 binary predictors (McNeish & Stapleton, 2016a). Type of institution (public versus private) is an example of a common level-2 predictor within the field of education (Peugh & Enders, 2005). Binary predictors with relatively constant prevalence between groups (e.g., treatment vs. control with 50:50 prevalence) function similarly as continuous predictors (McNeish & Stapleton, 2016b).

However, *SE* estimates will exhibit more bias when the prevalence of a binary predictor is "highly discrepant" (e.g., 10:90), especially when said predictor is included in an interaction (McNeish & Stapleton, 2016b, p. 302). A recent simulation showed *SE* estimates were biased until 60 level-2 clusters were obtained when a highly discrepant or unbalanced predictor (20:80) was part of an interaction (Bell, Schoeneberger, Smiley, Ene, & Leighton, 2013 as cited in McNeish & Stapleton, 2016b). Unfortunately, Lai et al. (2018) did not examine the efficiency of their proposed FPC adjustment method for *SE* estimates of binary predictors.

Results from Lai et al. (2018) demonstrate how the degree of bias increases with larger cluster size. Recall from Table 2.3 that the largest cluster size in their simulated conditions was 25. This presents a problem for applied researchers because cluster sizes of at least 30 (e.g., 30 students per classroom) are normal in educational research (Mass & Hox, 2005; McNeish & Stapleton, 2016a). Average cluster size ranges from 56 (Oberle et al., 2011) to 132 (Maxwell et al., 2017) level-1 units in the examples of multilevel studies with finite populations discussed above. Lai et al. (2018) purports their adjustment will produce acceptable *SE*s for larger cluster sizes (i.e., $\overline{n_j} > 25$) when the number of clusters is at least 30, but they did not actually test this claim and call for further research to "verify the performance of the adjusted *SE*s with larger cluster sizes" (p. 108).

The estimation of *SE*s in HLM analyses with few clusters is problematic for HLM analyses regardless of whether the target population of interest is considered finite or not (Mass & Hox, 2005; Snijders & Bosker, 2012). The FPC adjustment is susceptible to problems caused by a few number of clusters and tends to overcorrect estimates, resulting

in negatively biased *SE*s (Lai et al., 2018). Because of its limitations in small samples, Lai et al. (2018) suggest their adjustment should only be applied when the number of clusters is at least 30. Unfortunately, many of the samples utilized in education contain fewer than 30 level-2 units, as evident in the examples discussed above. Oberle et al. (2011) examined 25 level-2 units; Maxwell et al. (2017) examined 17 level-2 units; and Juvonen et al. (2011) examined 11 level-2 units. All of these examples include target populations that may be considered finite and fall short of Lai et al.'s (2018) suggestion of 30 level-2 units. Lai et al. (2018) concludes that for studies with small sample sizes (i.e., $J < 30$) "resampling techniques such as the bootstrap procedure in multilevel settings may be modified to accommodate the finite population and provide more robust standard error estimates" (p. 108). The following section introduces the bootstrap procedure and discusses how to incorporate an FPC adjustment into the bootstrap procedure for samples from finite level-2 populations.

Bootstrapping

The estimation of standard errors is problematic when dealing with a small number of clusters because traditional HLM analyses in those samples yield *SE* estimates that are too small (Dedrick et al., 2009; Mass & Hox, 2004; McNeish & Stapleton, 2016b; Snijders & Bosker, 2012). Bootstrapping methods have been presented as one option to deal with biased standard errors resulting from HLM analyses based on a few number of clusters (Dedrick et al., 2009; Lai et al., 2018; Mass & Hox, 2004; McNeish & Stapleton, 2016b; Snijders & Bosker, 2012).

Bootstrapping methods are part of the broad class of statistics commonly referred to as resampling methods (Chernick, 1999; Chernick & LaBudde, 2011). Note numerous

resampling methods exist (e.g., nonparametric bootstrap, parametric bootstrap, parametric residual bootstrap, jackknife, and delete-1 jackknife procedures (Goldstein, 2011; van der Leeden et al., 2007).) The method presented below is originally described in Efron (1979) and is generally referred to as Efron's nonparametric bootstrap.

Consider a scenario in which a random sample size of *n* is observed from an unspecified probability distribution *F*,

$$X_i = x_i,$$

where $X_i \sim iid\ F$, and *i* = 1, 2,…, *n* (Efron, 1979; Efron, 1982). $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ denotes the random sample and $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ denotes its observed realization. Let θ be some parameter of interest of *F* and $\hat{\theta}$ be an estimator of θ. The basic element for bootstrapping is the empirical distribution function of the observed data (Chernick & LaBudde, 2011). As such, the procedure assesses the accuracy of $\hat{\theta}$ in terms of its empirical distribution, F_n (Chernick, 1999). This empirical distribution function assigns equal probabilities of inclusion (i.e., $\pi_i = 1/n$) to each observed value x_i sampled. The bootstrap distribution (F_n*) for $\hat{\theta} - \theta$ is the distribution obtained by sampling independently with replacement from F_n.

There are n^n distinct bootstrap samples possible so practical application bootstrapping methods often requires a Monte Carlo approximation of the bootstrap estimate (Chernick & LaBudde, 2011). For this reason, bootstrapping methods are usually computer-based (Efron, 1993). In general, Efron's (1979) bootstrap procedure is implemented using the following steps:

1. Draw a sample size of *n* (where *n* is the original sample size) with replacement from the empirical distribution. Call this the bootstrap sample.
2. Compute θ*, the value of $\hat{\theta}$ obtained by using the bootstrap sample in place of the original sample.

3. Repeat steps 1 and 2 *B* times. (Chernick, 1999, p. 8).

The objective of Efron's (1979) bootstrapping method is to estimate a parameter θ based on the data without having to introduce parametric assumptions about the population distribution (Chernick & Labudde, 2011). θ can be any parameter of interest such as the mean, correlation, or standard deviation of *F* (Efron, 1979). As a result, the bootstrap procedure may be applied to any statistic (Efron & Gong, 1983).

There are two sources of error associated with the bootstrap procedure:

1. The Monte Carlo approximation to the bootstrap distribution
2. The approximation of the bootstrap distribution (F_n*) to the population distribution *F*. (Chernick & LaBudde, 2011, p. 5).

Errors associated with the Monte Carlo approximation to the bootstrap distribution are minimized by increasing the number of bootstrap replications *B*. The bootstrapping procedure "works" if F_n converges to *F* as $n \rightarrow \infty$ (Chernick & LaBudde, 2011, p. 5).

The bootstrap procedure described in this section is considered a nonparametric resampling method because no parametric assumptions about the population distribution are introduced. The method described literally resamples from the empirical distribution of the observed data F_n (Davison & Hinkley, 1997). F_n is the maximum likelihood estimator of the distribution of the observations when no parametric assumptions are made (Chernick, 1999). Although no *parametric assumptions* are made, this does not mean that the nonparametric bootstrap procedure is *assumption free*. The nonparametric bootstrapping procedure described assumes $X_i \sim iid\ F$ (Efron, 1979; Efron, 1982). Consequently, Efron's (1979) nonparametric bootstrapping procedure described in this section cannot be applied to hierarchical data structures, such as those described above in the examples of multilevel studies dealing with finite populations, because the

independence of observations at level-1 is conditional on the level-2 units (van der Leeden et al., 2007).

Multilevel Bootstrapping

Resampling schemes for hierarchical data structures must account for the fact that observations are subject to intra-class dependency (van der Leeden et al., 2007). Hence, bootstrap resampling methods for HLMs must account for the dependency of observations due informative sampling designs (Goldstein, 2011). This section summarizes two multilevel bootstrapping strategies that retain nested data structures. Both strategies are considered multilevel extensions of Efron's (1979) nonparametric bootstrapping procedure.

Resampling complete level-2 units. The first strategy for retaining nested data structures keeps the selected level-2 units intact and is implemented using the following steps:

1. Draw a sample of size J with replacement from the level-2 units.
2. For each j, draw n_j cases without replacement from the level-2 units selected in step 1 (i.e., select all the level-1 units for each level-2 unit selected).
3. Compute estimates for parameters of the two-level model.
4. Repeat steps 1-3 B times. (Davison & Hinkley, 1997, p. 100).

Resampling level-1 units within level-2. Another strategy for retaining nested data structures involves resampling level-1 units within resampled level-2 units and is implemented using the following steps:

1. Draw a sample of size J with replacement from the level-2 units.
2. For each j, draw n_j cases with replacement from the level-2 units selected in step 1 (i.e., resample from all the level-1 units for each level-2 unit selected)
3. Compute estimates for parameters of the two-level model.
4. Repeat steps 1-3 B times. (Davison & Hinkley, 1997, p. 100).

Finite Population Bootstrapping

Efron's (1979) nonparametric bootstrapping procedure and its multilevel extension both involve resampling with replacement samples of original size and do not capture the effect of the sampling fraction. For single level studies, Davison and Hinkley (1997) suggest resampling with replacement samples of size

$$n' = (n-1)/(1-f) \tag{2.30}$$

in order "to shrink the variance of an estimator" and deal with finite populations (p. 93). Extending this logic to the HLM framework suggests that a plausible strategy for dealing with finite populations at level-2 is to resample with replacement samples of size

$$J' = (J-1)/(1- f_1)\ , \tag{2.31}$$

where f_1 is calculated according to Equation 2.20.

The finite population bootstrapping procedure for hierarchical data structures with two-levels utilized in the current study is implemented using the following steps:

1. Draw a sample of size *J'* with replacement from the level-2 units.
2. For each *j*, draw n_j cases without replacement from the level-2 units selected in step 1 (i.e., select all the level-1 units for each level-2 unit selected).
3. Compute estimates for parameters of the two-level model.
4. Repeat steps 1-3 *B* times.

Lai et al. (2018) "assumed that the level-1 variables were sampled from an infinite population" to simplify their simulation because it is more common to "make inference to a finite level-2 population" (p. 102). Therefore, the finite population bootstrapping procedure chosen for the current study resamples complete level-2 units only. Furthermore, only resampling complete level-2 units requires a lighter computational load than resampling level-1 units within level-2.

Summary

Applied researchers need to carefully define their populations of interest and consider the characteristics of their samples because appropriate modes of statistical inference depend on the properties of their sampling designs. Some sampling designs, especially those used in educational settings, result in hierarchical data structures. HLMs are needed to appropriately analyze hierarchical data structures and are considered an integrated approach because they model data while accounting for important features of the sampling design, that is, clustering.

The theory behind HLM was developed for cases where observations are sampled from an infinite superpopulation. Sometimes the level-2 units are few and countable (i.e., finite) as evident in the empirical examples discussed above. Using HLM when the level-2 sample size exceeds the finite population by as little as 5% results in overestimated standard errors and confidence intervals that are too wide.

Lai et al. (2018) proposed an FPC adjustment method for fixed effect standard errors for two-level HLMs (described above) and evaluated its performance using Monte Carlo simulations. Lai et al.'s (2018) approach integrated an additional design feature (i.e., the sampling fraction *f*) into the traditional HLM framework and produced unbiased standard errors when the number of level-2 units was at least 30. However, studies based on fewer than 30 level-2 units are common in educational settings, as in the examples of multilevel studies in which the population is finite discussed above. Lai et al. (2018) suggested future research evaluate the performance of their proposed adjustment against bootstrapping procedures when the number of level-2 units is fewer than 30.

Lai et al. (2018) leaves many questions for applied researchers considering applying FPCs to their hierarchical data, especially when dealing with a few number of large clusters. It remains unknown whether Lai et al.'s (2018) FPC adjustment produces acceptable standard errors when the number of level-2 units is fewer than 30 or when the number of level-1 units is greater than 25, both of which are common in educational settings. Also, it remains unknown how Lai et al.'s (2018) FPC adjustment compares to bootstrapping alternatives. Finally, it remains unknown whether Lai et al.'s (2018) FPC adjustment produces acceptable standard errors for binary predictors.

Current Study

The purpose of the current study is to evaluate Lai et al.'s (2018) FPC adjustment in two-level hierarchical linear models for a few number of large clusters, compare the FPC adjustment's performance to a finite population bootstrapping alternative, and examine the efficiency the FPC adjustment for standard errors associated with a binary level-2 predictor. Monte Carlo simulation methods were used to assess the effects of the following factors: (a) number of clusters in the sample (20, 30, and 60); (b) cluster size (30, 90, and 150); (c) analysis method (no bootstrap unadjusted, no bootstrap FPC adjusted, and finite population bootstrap); (d) binary predictor ratio (50:50 and 20:80); and (e) binary predictor effect (γ_{02} = .45 and γ_{02} = .20). Recall, the two factors relating to the binary predictor were isolated in their own study because of anticipated convergence issues for models using binary predictors. Refer to Tables 1.1 and 1.2 for summaries of the simulation design factors and their levels for each study. Specific research questions for the continuous predictors study were:

RQ1a. How do unadjusted standard errors (SE_0s) compare to FPC adjusted standard errors (SE^{FP}s)?

RQ1b. How do SE_0s compare to SE^{FP}s for data with a few number of large clusters (i.e., $J < 30$ & $n_j > 25$)?

RQ2. How do finite population bootstrapped standard errors (SE^{FPboot}s) compare to SE^{FP}s?

The specific research questions for the binary predictor study were

RQ3. How do SE_0s compare to SE^{FP}s for binary predictors?

RQ4. How do SE^{FPboot}s compare to SE^{FP}s binary predictors?

The current study adds to the body of literature in the following ways. First, it tailors the use of Lai et al.'s (2018) FPC adjustment method specifically to sample sizes common in educational settings. Second, it compares the FPC adjustment to finite population bootstrapping alternatives. Finally, it evaluates the efficiency of the FPC adjustment for the *SE*s of binary predictors.

CHAPTER III. METHODS AND PROCEDURES

A Monte Carlo simulation study was conducted to evaluate the performance of Lai et al.'s (2018) FPC adjustment method across the design factors listed in Tables 1.1 and 1.2. This chapter begins with a discussion of the models used to generate the data. Following that, the simulation conditions and their rationale are described. Finally, the criteria used to evaluate the research questions above are explained.

Data Generation

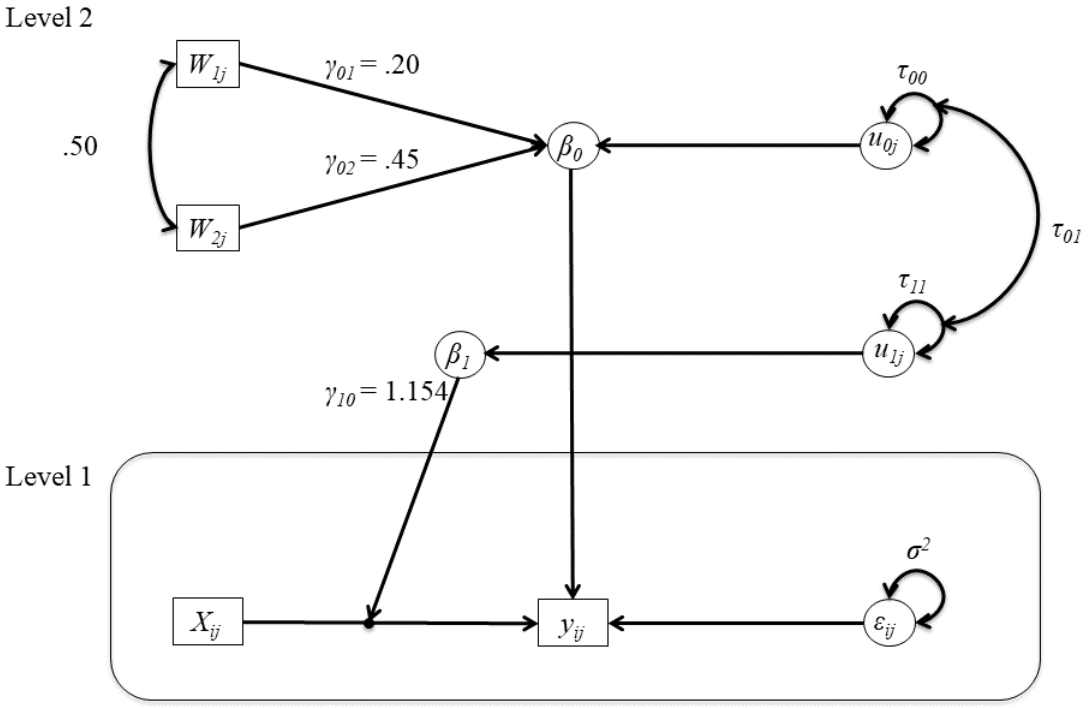
Continuous Predictors Study

Data for the current continuous predictors study were generated according to the following random slope and intercept model utilized by Lai et al. (2018):

$$y_{ij} = \gamma_{00} + \gamma_{01}W_{1j} + \gamma_{02}W_{2j} + \gamma_{10}X_{ij} + u_{1j}X_{ij} + u_{0j} + \varepsilon_{ij} \quad (3.1)$$

(p. 101). The data generating model had two level-2 predictors (W_{1j} and W_{2j}) and one level-1 predictor (X_{ij}) as shown in Figure 3.1.

Figure 3.1. Path diagram for data-generating model.



The parameters used to generate the data for the continuous predictors study were identical to the parameters used in Lai et al. (2018): The intercept γ_{00} was set to 0. Both W_{1j} and W_{2j} were normally distributed with a mean of 0 and variance of 1. The correlation was set to $r_{W_{1j}W_{2j}} = .5$. X_{ij} was normally distributed with a mean of 2 and variance of 1. u_{1j} followed a normal distribution with a mean of 0 and a variance $\tau_{11} = .5$. γ_{01} and γ_{02}were set as .2 and .45 to represent small and medium effects, respectively. Both u_{0j} and ε_{ij} were normally distributed with a mean of 0. The variances of u_{0j} and ε_{ij}were set to $\tau_{00} = 1$ and $\sigma^2 = 4$, respectively. γ_{10} was set to 1.154. y_{ij} was computed according to Equation 3.1.

Binary Predictor Study

Data for the current binary predictor study were generated according to the following model:

$$y_{ij} = \gamma_{00} + \gamma_{01}W_{1j} + \gamma_{02}W_{2j}^{B} + \gamma_{10}X_{ij} + u_{1j}X_{ij} + u_{0j} + \varepsilon_{ij} \; . \qquad (3.2)$$

Data generation for the binary predictor study was identical to the continuous predictors study, except W_{2j}^{B} was Bernoulli distributed with a probability of success determined by the binary predictor ratio and the level-2 fixed effects (i.e., γ_{01}=.20 and $\gamma_{02} = .45$ in Figure 3.1) were transposed when the binary predictor effect $\gamma_{02} = .20$. Please see the following sections on binary predictor ratio and effect.

Generating Data Using Copulas

Lai et al. (2018) generated their level-2 variables (i.e., W_{1j}, W_{2j}, u_{0j}, and u_{1j}) from a multivariate normal distribution with the covariance matrix

$$\begin{bmatrix} 1 & .50 & 0 & 0 \\ .50 & 1 & 0 & 0 \\ 0 & 0 & 1 & .25 \\ 0 & 0 & .25 & .50 \end{bmatrix}$$

(p. 102). However, generating data from a multivariate normal distribution is not an appropriate strategy for the binary predictor study because W_{2j}^{B} is Bernoulli distributed, rather than normal. Because of this issue, data for both the continuous predictors and binary predictor studies were generated using copulas.

Copulas have become a popular tool in situations where multivariate normality is questionable (Yan, 2007). A copula is a function that joins or "couples" multivariate distribution functions to their one-dimensional marginal distribution functions (Nelson,

1999). More specifically, a copula is "merely a *d*-dimensional cumulative distribution function with standard uniform margins" (Kojadinovic & Yan, 2010, p. 1).

One of the primary applications of copulas is in Monte Carlo simulations (Nelson, 1999). Data for both the continuous predictors and binary predictor studies were generated using copulas because of the combination of continuous and a binary predictor in the binary predictor study. Specifically, data for the level-2 variables were generated from a 4-dimensinal copula to preserve the covariance matrix presented above using the copula package in R (Hofert, Kojadinovic, Maechler, & Yan, 2018).

Simulation Conditions & Their Justification

Study conditions were chosen based on their importance in past research on sufficient sample sizes for multilevel modeling, their use in other multilevel simulation studies, and their prevalence in applied educational settings.

Continuous Predictors Study

Number of clusters in the sample. The degree of bias in SE_0s for level-2 effects increases with the number of clusters (Lai et al., 2018). Lai et al. (2018) suggest their FPC adjustment should only be applied when the number of clusters is at least 30. However, the samples used in educational settings may contain fewer than 30 clusters, as discussed in the examples of multilevel studies above. A recent simulation demonstrated how *SE* estimates were biased until 60 level-2 clusters were obtained when an unbalanced binary predictor was included the model (Bell, Schoeneberger, Smiley, Ene, & Leighton, 2013 as cited in McNeish & Stapleton, 2016b). Accordingly, clusters of 20, 30, and 60 were chosen for the current study.

Cluster size. The degree of bias in SE_0s for level-2 effects also increases with cluster size, but to a smaller degree than with number of clusters (Lai et al., 2018). Twenty-five was largest average cluster size examined in Lai et al. (2018). However, cluster sizes of at least 30 are normal in educational research (Mass & Hox, 2005; McNeish & Stapleton, 2016a). For example, Thrash and Warner (2016) examined an average of 93 students per school. Maxwell et al. (2017) examined an average of 133 students per school. Cluster sizes of 30, 90, and 150 were chosen for the current study to correspond with the number of level-1 units used in the examples of multilevel studies in educational settings.

Analysis method. Lai et al. (2018) concludes "resampling techniques such as the bootstrap procedure in multilevel settings may be modified to accommodate the finite population and provide more robust standard error estimates" and encourages future researchers to implement a finite population bootstrap procedure and "evaluate its performance against" their proposed FPC adjustment (p. 108). The current study provides that evaluation by comparing standard errors from the following analysis methods: no bootstrap, unadjusted (SE_0s); no bootstrap, FPC adjusted (SE^{FP}s); and finite population bootstrap (SE^{FPboot}s).

Unique Conditions for Binary Predictor Study

The factors relating to the binary predictor were isolated in their own study because of anticipated convergence issues for models using binary predictors.

Cluster size. The bias in SE_0s for level-2 effects increases with cluster size, but to a smaller degree than with number of clusters (Lai et al., 2018). Lai et al. (2018) "assumed that the level-1 variables were sampled from an infinite population" to simplify

their simulation because it is more common to "make inference to a finite level-2 population" (p. 102). Following the same logic, manipulating level-1 cluster size was a not primary concern so n_j was set to 30 for the binary predictor study.

Binary predictor ratio. Binary predictors with a relatively constant prevalence between groups function similarly as continuous predictors (McNeish & Stapleton, 2016b). However, *SE* estimates will exhibit more bias when the prevalence of a binary predictor is "highly discrepant" or unbalanced (McNeish & Stapleton, 2016b). For the binary predictor study, W_{2j}^{B} in Equation 3.2 was Bernoulli distributed with a probability of success (i.e., probably of being coded 1) determined by the binary predictor ratio. Ratios of 50:50 and 20:80 were used to represent relatively constant or balanced and discrepant or unbalanced binary predictors, respectively.

Binary predictor effect. The level-2 effect size for the binary predictor (γ_{02}) had two levels. For the first level, the effect of the continuous predictor was γ_{01}=.20 and the effect of the binary predictor was $\gamma_{02} = .45$ as in Figure 3.1. For the second level, the level-2 effects were transposed such that the effect of the continuous predictor was γ_{01}=.45 and the effect of the binary predictor was $\gamma_{02} = .20$.

Constants

The following factors were held constant across all conditions in both the continuous predictors and binary predictor studies.

Sample-population ratio. The bias in SE_0s for level-2 effects increases with sample-population ratio (*P*) and *P* explains the most variability in bias relative to the other factors manipulated in Lai et al. (2018). Refer to Table 2.3 for a summary of design factors from Lai et al. (2018). This association between bias and *P* is well documented in

the literature. For example, Cochran's 1977 text discusses how FPCs need to be applied to correct biased sample estimates when *P* is as little as 5%. Manipulating the sample-population ratio in the current study was not expected to contribute to the established body of literature. Consequently, *P* = .25 across all conditions listed in Tables 1.1 and 1.2.

ICC. The degree of bias in SE_0s for level-2 effects also increases with higher ICC (ρ) (Lai et al., 2018). An ICC of zero indicates no variation in the outcome of interest across level-2 units (i.e., no dependency) and no FPC for level-2 is needed whether the level-2 population is considered finite or infinite. The ICC's effect on standard error estimates is well documented in the literature (e.g., Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). Because of its established effect in the literature, ρ was not manipulated in the current study.

ρ was calculated according to Equation 2.6. Recall, the variance of u_{0j} and ε_{ij}were set to $\tau_{00} = 1$ and $\sigma^2 = 4$, respectively. Substituting those values into Equation 2.6 yields an ρ = .20. Consequently, ρ = .20 across all conditions in both the continuous predictors and binary predictor studies.

Procedure

All data were generated in R 3.6.0 (R Core Team, 2019) via the RStudio 1.1.463 interface (RStudio Team, 2018). See Appendix B for code used to conduct the simulation. Five hundred finite populations (i.e., population replications) the size of $J_{pop} = J/P$ were generated for each condition to ensure the results did not depend on the characteristics of a single finite population. There were 500 samples (i.e., sample replications) size of *J* complete level-2 units drawn without replacement from each

population. Then, there were 500 bootstrapped samples (i.e., bootstrap replications) size of J' (rounded up to the nearest integer) complete level-2 units (see Equation 2.31) drawn with replacement from each sample. Thus, a total of

$500 + 500^2 + 500^3 = 125{,}250{,}500$ data sets were created for each condition.

The data generating model (i.e., Equation 3.1) was fit to each sample and bootstrapped sample data sets using restricted maximum likelihood estimation using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015). Empirical standard errors were obtained for each population. Estimated fixed effects at level-1 (i.e., $\widehat{\gamma_{10}}$), at level-2 (i.e., $\widehat{\gamma_{01}}$ and $\widehat{\gamma_{02}}$), and their unadjusted standard errors (SE_0s) were obtained for each sample data set. Then, FPC adjusted standard errors (SE^{FP}s) were computed for each $\hat{\gamma}$ by taking the square root of the diagonal elements of $\mathrm{Var}^{\mathrm{FP}}(\hat{\boldsymbol{\gamma}})$ from each sample data set (see Equation 2.29). Finally, the finite population bootstrapping procedure was implemented and finite population bootstrapped standard errors (SE^{FPboot}s) for the $\hat{\gamma}s$ were obtained for each sample data set. SE^{FPboot}s were averaged across the 500 bootstrapped replications for each sample.

Computational Intensity Pilot Study

A preliminary study was conducted to estimate the computational time required for the simulation due to the large number of data sets analyzed for each condition. To be conservative, the largest sample size condition from the continuous predictors study (i.e., $J = 60$ and $n_j = 150$) was used for the computational intensity pilot. Data were generated and analyzed according to the procedure above for only the first 50 complete replications (i.e., 50 population replications X 50 sample replications X 50 bootstrap

replications) or .001 of the total condition. The package tictoc (Izrailev, 2014) was used to record computational time.

The computational intensity pilot took 2.822497 hours (10,160.99 seconds) to complete on a 64-bit operating system with 16 GB of RAM and a 3.50 GHz processor. This suggested a single condition may require 117.604 days to complete. The complete current study required running 21 simulation conditions (analysis method factors were analyzed simultaneously). Recall Tables 1.1 and 1.2.

The computational intensity pilot suggested the full simulation may take 6.766258 years to complete on a conventional operating system. Consequently, the current study was conducted on Crane at the UNL Holland Computing Center.

Evaluation Criteria

SE_0, SE^{FP}, and SE^{FPboot} were evaluated for each population using relative bias, mean squared error (MSE), and root mean square error (RMSE). Coverage was an additional criterion used to evaluate the binary predictor.

Relative Bias

SE_0, SE^{FP}, and SE^{FPboot} from each sample *i* were compared to the empirical standard errors (SD_j) of each $\widehat{\boldsymbol{\gamma}}$ for the *j*th population such that

$$Relative\ Bias\left[SE_j(\hat{\gamma})\right] = \frac{[\sum_i SE_j(\hat{\gamma}_i)]/R - SD_j(\hat{\gamma})}{SD_j(\hat{\gamma})}, \tag{3.3}$$

where *R* is the number of sample replications (i.e., *R* = 500). The empirical standard error is the standard deviation of a given parameter's estimates across replications or repeated samples,

$$SD_j(\hat{\gamma}) = \sqrt{\frac{\sum_i(\hat{\gamma}_i - \bar{\hat{\gamma}}_i)^2}{(R-1)}}, \tag{3.4}$$

where $\overline{\widehat{\gamma}_i}$ is the mean of the estimate of $\boldsymbol{\gamma}$ across sample replications (Hoogland & Boomsma, 1998). Relative bias was averaged across the 500 finite populations for each condition.

Mean Square Error

The mean square error (MSE) was computed for each population such that

$$MSE\left[SE_j(\hat{\gamma})\right] = \frac{\sum_i [SE_j(\hat{\gamma}_i) - SD_j(\hat{\boldsymbol{\gamma}})]^2}{R}. \quad (3.5)$$

As was done with relative bias, MSE was averaged across the 500 finite populations for each condition.

Root Mean Square Error

The RMSE for each condition was calculated by taking the square root of the averaged MSEs.

Binary Predictor Coverage

The estimate of the binary predictor's effect (i.e., γ_{02} for the binary predictor study only) was evaluated in terms of coverage. Coverage of the population parameter was calculated as

$$Coverage_{\boldsymbol{\gamma}} = \frac{\sum_i I\ (\boldsymbol{\gamma} \in (\hat{\gamma}_i \pm 1.96 \times SE(\hat{\gamma}_i)))}{R}, \quad (3.6)$$

where *I* is an indicator function that takes on a value of 1 if the interval estimate for sample *i* contains the population parameter ($\boldsymbol{\gamma}$), and a 0 otherwise. Whether the interval estimate of the binary predictor contains zero ($Coverage_0$) was also of interest because binary predictor effect was a manipulated condition in the current binary predictor study. $Coverage_0$ was calculated as

$$Coverage_0 = \frac{\sum_i I\ (0 \in (\hat{\gamma}_i \pm 1.96 \times SE(\hat{\gamma}_i)))}{R}, \quad (3.7)$$

where I is an indicator function that takes on a value of 1 if the interval estimate for sample i contains zero, and a 0 otherwise.

CHAPTER IV. RESULTS

Continuous Predictors Study

Results associated with the unadjusted standard errors (SE_0) and the FPC adjustment (SE^{FP}) from the continuous predictors study coincided with those from Lai et al. (2018). The average relative biases of each predictor's effects are displayed in Table 4.1. As shown, the unadjusted standard error estimates (SE_0) overestimated the empirical *SE*s, whereas the FPC adjustment estimates (SE^{FP}) and the finite population bootstrap estimates (SE^{FPboot}) underestimated the empirical *SE*s. The relative bias of SE^{FP} was closer to zero than the relative bias of SE_0. Each level-2 effect and the level-1 effect for the continuous predictors study are discussed in turn in the following sections.

Table 4.1. *Average Relative Bias Across Conditions for Continuous Predictors Study*

Effect	SE_0	SE^{FP}	SE^{FPboot}
γ_{01}	.084	-.061	-.217
γ_{02}	.077	-.067	-.219
γ_{10}	.151	-.003	-.296

Note. SE_0 = unadjusted standard error estimates; SE^{FP} = FPC adjustment estimates; SE^{FPboot} = finite population bootstrap estimates.

Level-2 Effects

γ_{01}. Factors explaining the relative bias of unadjusted standard error estimates of γ_{01} are displayed in Table 4.2. The number of clusters (*J*) was the effect most associated with variability in relative bias for SE_0 of γ_{01}.

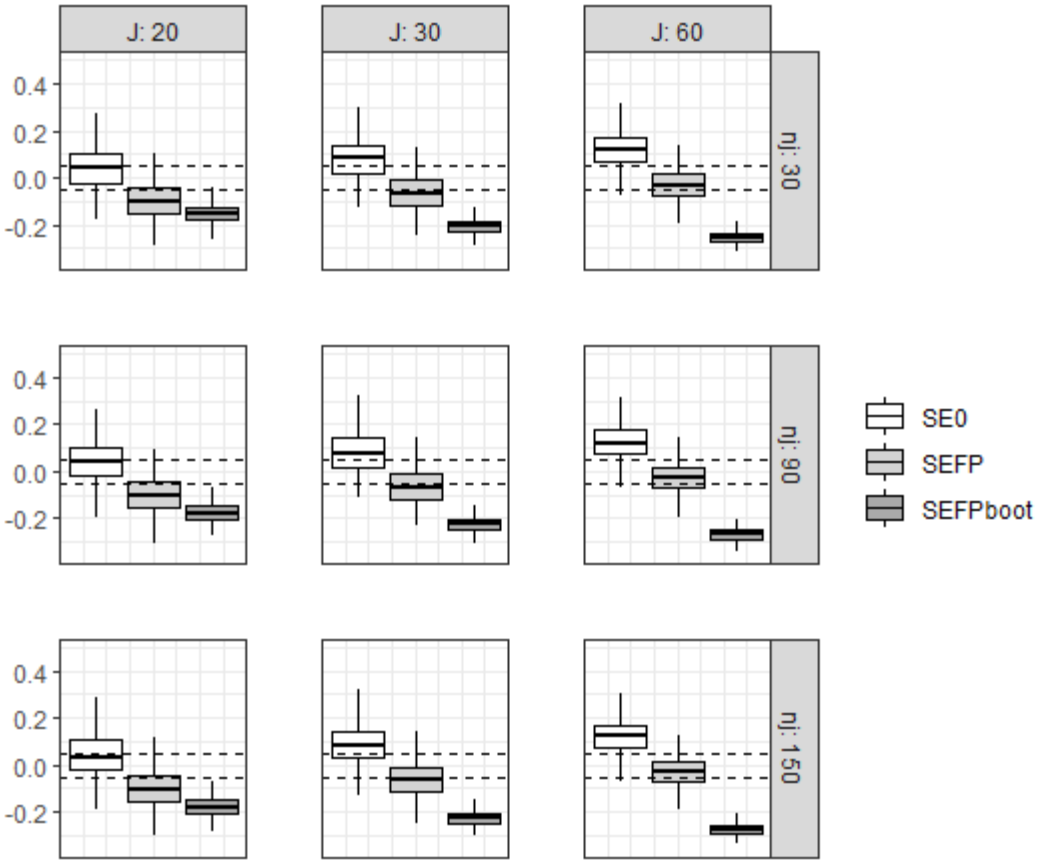
Table 4.2. *ANOVA for Relative Bias in SE_0 of γ_{01} for Continuous Predictors Study*

Effect	Sum Sqr	Df	Mean Sqr	*F*	*p*	Partial η^2
J	4.793	2	2.397	319.708	<.001	.125
n_j	.012	2	.006	.773	.462	<.001
J x n_j	.018	4	.005	.609	.656	.001

Note. ANOVA = analysis of variance; SE_0 = unadjusted standard error estimates; *J* = number of clusters in sample; n_j = number of elements drawn from each cluster.

The relative biases for γ_{01} are compared in Figure 4.1, with boxplots displaying the average SE_0, SE^{FP}, and SE^{FPboot} estimates across populations for each condition and dashed lines representing acceptable levels according to Hoogland and Boomsma's (1998) guidelines (i.e., $|Relative\ Bias| < .05$). The degree of relative bias in the unadjusted standard error estimates increased with larger J and larger n_j as shown in Figure 4.1. Relative bias in SE_0s exceeded .05 and was non-ignorable when $J > 20$.

Figure 4.1. Percentage relative bias in *SE*s for γ_{01} in the continuous predictors study.



The FPC adjusted standard error estimates exhibited acceptable levels of relative bias with larger J and larger n_j. However, SE^{FP}s were negatively biased (i.e., the adjustment underestimated the empirical *SE*s) when $J = 20$ as shown in Figure 4.1.

The finite population bootstrap tended to underestimate the empirical *SE*s across all conditions. The degree of relative bias in SE^{FPboot}s increased with larger *J* and larger n_j as shown in Figure 4.1.

MSE and RMSE for γ_{01} are presented in Table 4.3. As shown, the SE^{FP} estimates were closer to the empirical *SE*s than the SE_0 estimates across all conditions and the SE^{FPboot} estimates exhibited the most error. The amount of error in estimates decreased as the sample size (i.e., *J* and n_j) increased.

Table 4.3. *Mean Square Error and Root Mean Square Error for γ_{01} in Continuous Predictors Study*

		MSE (RMSE)		
J	n_j	SE_0	SE^{FP}	SE^{FPboot}
20	30	.0078 (.0872)	.0072 (.0834)	.0100 (.0983)
	90	.0054 (.0727)	.0051 (.0697)	.0078 (.0869)
	150	.0050 (.0699)	.0047 (.0670)	.0073 (.0844)
30	30	.0036 (.0587)	.0028 (.0519)	.0055 (.0729)
	90	.0024 (.0479)	.0018 (.0421)	.0042 (.0642)
	150	.0023 (.0466)	.0017 (.0407)	.0040 (.0624)
60	30	.0011 (.0326)	.0006 (.0240)	.0025 (.0469)
	90	.0008 (.0276)	.0004 (.0198)	.0019 (.0437)
	150	.0007 (.0260)	.0004 (.0189)	.0019 (.0427)

Note. MSE = mean square error; RMSE = root mean square error; *J* = number of clusters in sample; n_j = number of elements drawn from each cluster; SE_0 = unadjusted standard error estimates; SE^{FP} = FPC adjustment estimates; SE^{FPboot} = finite population bootstrap estimates. Values were rounded to the nearest .0001.

γ_{02}. Average relative bias of γ_{02} was similar to γ_{01} (i.e., within .01). Refer to Table. 4.1. Consequently, the interpretation of the results for $\gamma_{02}$was identical to the interpretation for γ_{01}. Factors explaining the relative bias of unadjusted standard error estimates of γ_{02} are displayed in Table 4.4. Similar to the other level-2 effect (i.e., γ_{01}), *J* was the effect most strongly associated with the variability in relative bas for SE_0s of γ_{02}.

Table 4.4. *ANOVA for Relative Biases in SE_0 of γ_{02} for Continuous Predictors Study*

Effect	Sum Sqr	Df	Mean Sqr	*F*	*p*	Partial η^2
J	5.562	2	2.781	364.827	<.001	.140
n_j	.012	2	.006	.759	.468	<.001
J x n_j	.008	4	.002	.264	.901	<.001

Note. ANOVA = analysis of variance; SE_0 = unadjusted standard error estimates; J = number of clusters in sample; n_j = number of elements drawn from each cluster.

Figure 4.2 shows the percentage relative bias for γ_{02}. The degree of relative bias in SE_0estimates increased with larger sample size. SE^{FP}s were negatively biased when J = 20, and SE^{FPboot}s underestimated the empirical *SE*s across all conditions.

Figure 4.2. Percentage relative bias in *SE*s for γ_{02} in the continuous predictors study.

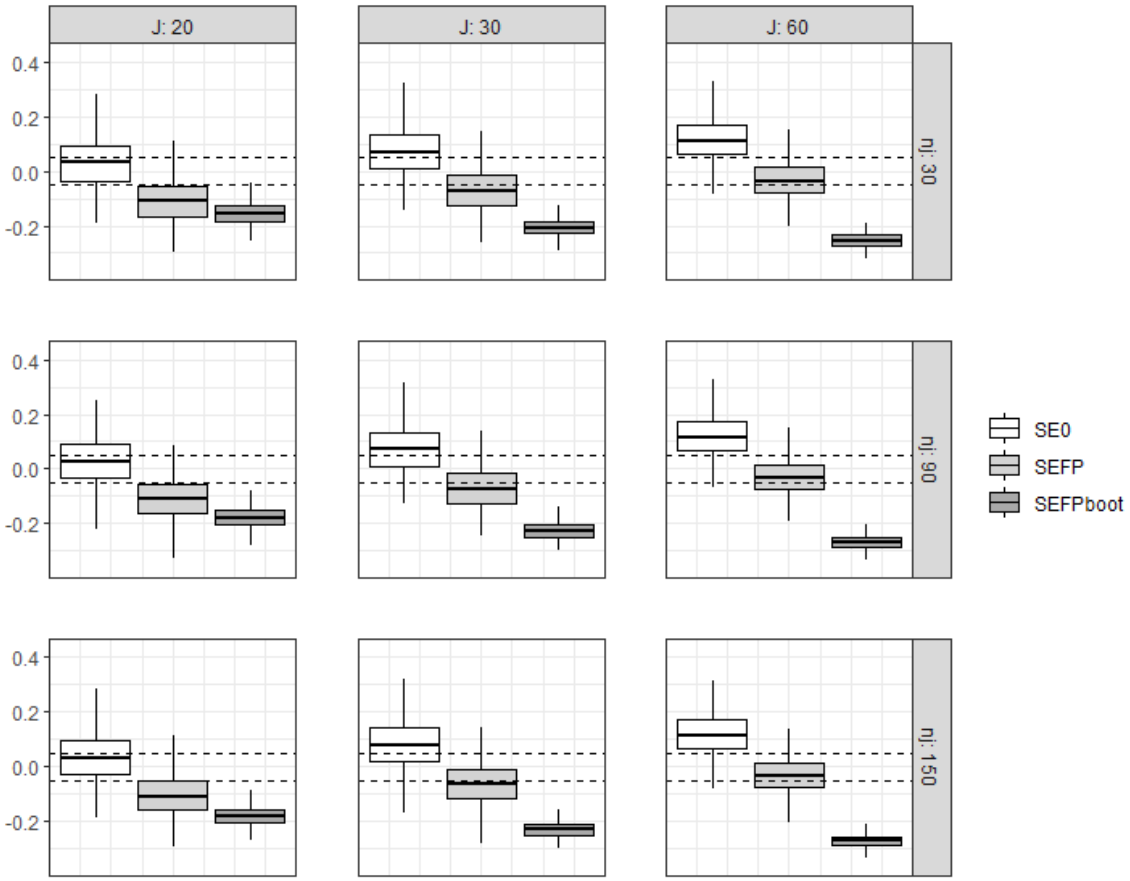


Table 4.5 displays the MSE and RMSE for γ_{02}. As shown, the SE^{FP} estimates were closer to the empirical *SE*s than the SE_0s, and the SE^{FPboot}s exhibited the most error across all conditions.

Table 4.5. *Mean Square Error and Root Mean Square Error for γ_{02} in Continuous Predictors Study*

		MSE (RMSE)		
J	n_j	SE_0	SE^{FP}	SE^{FPboot}
20	30	.0077 (.0866)	.0074 (.0845)	.0102 (.0994)
	90	.0055 (.0733)	.0054 (.0719)	.0081 (.0887)
	150	.0048 (.0678)	.0046 (.0663)	.0073 (.0840)
30	30	.0035 (.0578)	.0028 (.0524)	.0055 (.0729)
	90	.0023 (.0475)	.0019 (.0431)	.0044 (.0654)
	150	.0022 (.0454)	.0017 (.0403)	.0040 (.0625)
60	30	.0011 (.0324)	.0006 (.0242)	.0025 (.0498)
	90	.0008 (.0272)	.0004 (.0200)	.0020 (.0443)
	150	.0007 (.0257)	.0004 (.0189)	.0018 (.0423)

Note. MSE = mean square error; RMSE = root mean square error; *J* = number of clusters in sample; n_j = number of elements drawn from each cluster; SE_0 = unadjusted standard error estimates; SE^{FP} = FPC adjustment estimates; SE^{FPboot} = finite population bootstrap estimates. Values were rounded to the nearest .0001.

Level-1 Effect

γ_{10}. Factors explaining the relative bias of unadjusted standard error estimates of γ_{10} are displayed in Table 4.6. Cluster size (n_j) and its interaction with number of clusters (i.e., *J* x n_j) explained more variability in unadjusted relative bias than *J* for the level-1 effect. The factor n_j was the effect most associated with variability in relative bias for SE_0s of the level-1 effect, whereas *J* was the effect most associated with variability in relative bias for SE_0s of the level-2 effects. See Tables 4.2, 4.4, and 4.6.

Table 4.6. *ANOVA for Relative Biases in SE_0 of γ_{10} for Continuous Predictors Study*

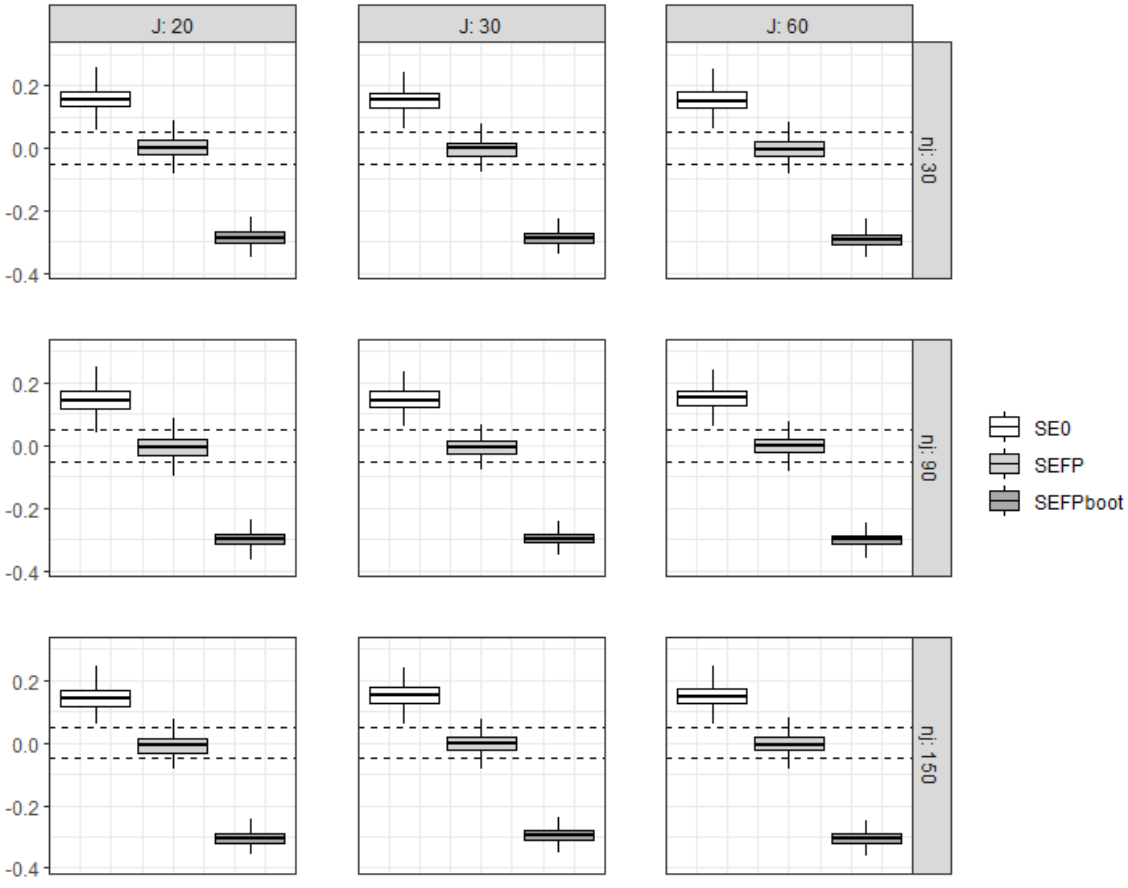
Effect	Sum Sqr	Df	Mean Sqr	*F*	*p*	Partial η^2
J	.009	2	.005	3.310	.037	.001
n_j	.022	2	.011	8.019	<.001	.004
J x n_j	.042	4	.010	7.526	<.001	.007

Note. ANOVA = analysis of variance; SE_0 = unadjusted standard error estimates; *J* = number of clusters in sample; n_j = number of elements drawn from each cluster.

There was substantial bias in SE_0s for the level-1 effect (i.e., γ_{10}) across conditions, but the FPC correction removed virtually all that bias. Refer to Table 4.1. The SE^{FPboot}s underestimated the empirical *SE*s for γ_{10}, similar to the level-2 effects.

The relative biases for γ_{01} are shown in Figure 4.3. The degree of relative bias in SE_0s was unacceptable because it exceeded Hoogland and Boomsma's (1998) guideline (i.e., $|Relative\ Bias| < .05$) regardless of J and n_j. The SE^{FP}s were within the acceptable range across all conditions. The SE^{FPboot}s underestimated the empirical *SE*s across all conditions.

Figure 4.3. Percentage relative bias in *SE*s for γ_{10} in the continuous predictors study.



MSE and RMSE for γ_{10} are presented in Table 4.7. As shown, the SE^{FP}s were closer to the empirical *SE*s than the SE_0s across all conditions. Moreover, the SE^{FPboot}s

exhibited the most error although the amount of error in the estimates decreased as the sample size increased.

Table 4.7. *Mean Square Error and Root Mean Square Error for γ_{10} in Continuous Predictors Study*

		MSE (RMSE)		
J	n_j	SE_0	SE^{FP}	SE^{FPboot}
20	30	.0013 (.0352)	.0005 (.0225)	.0022 (.0468)
	90	.0010 (.0313)	.0004 (.0206)	.0021 (.0456)
	150	.0009 (.0305)	.0004 (.0203)	.0020 (.0449)
30	30	.0007 (.0257)	.0002 (.0151)	.0015 (.0378)
	90	.0006 (.0233)	.0002 (.0138)	.0013 (.0363)
	150	.0005 (.0231)	.0002 (.0135)	.0013 (.0352)
60	30	.0003 (.0160)	.0001 (.0078)	.0007 (.0269)
	90	.0002 (.0148)	.0001 (.0071)	.0006 (.0253)
	150	.0002 (.0144)	.0001 (.0070)	.0006 (.0253)

Note. MSE = mean square error; RMSE = root mean square error; J = number of clusters in sample; n_j = number of elements drawn from each cluster; SE_0 = unadjusted standard error estimates; SE^{FP} = FPC adjustment estimates; SE^{FPboot} = finite population bootstrap estimates. Values were rounded to the nearest .0001.

Binary Predictor Study

As expected, many of the models did not converge when sampling from a finite population with a binary level-2 predictor (i.e., W_{2j}^{B} in Equation 3.2), resulting in missing model estimates (i.e., $\widehat{\gamma_{02}}$) from lme4. Table 4.8 shows the number of finite populations with complete data from their sample replications.

Table 4.8. *Number of Populations in Binary Predictor Study with Complete Sample Replications*

	Binary predictor ratio	
J	.2	.5
20	180 (18.0%)	1000 (100%)
30	708 (70.8%)	1000 (100%)
60	1000 (100%)	1000 (100%)

Note. J = number of clusters in sample.

As can be seen, non-convergence was an issue when drawing a few clusters from a population with an unbalanced binary predictor (i.e., binary predictor ratio = 20:80).

Only populations with complete sample replications (i.e., estimates of γ_{02}) were included in the current study. Consequently, 4,888 finite populations were analyzed in the binary predictor study.

Non-convergence issues were compounded by drawing bootstrapped samples from the random samples (i.e., none of the bootstrap replications' models converged if their sample replication's model failed). See Table 4.9.

Table 4.9. *Number of Populations in Binary Predictor Study with Complete Bootstrap Replications*

	Binary predictor ratio	
J	.2	.5
20	0 (00.0%)	2 (00.2%)
30	0 (00.0%)	860 (86.0%)
60	196 (19.6%)	1000 (100%)

Note. *J* = number of clusters in sample.

Non-convergence in the bootstrapped samples was almost certain as the number of clusters decreased and when the binary predictor was unbalanced as shown in Table 4.9. Consequently, finite population bootstrap estimates (SE^{FPboot}s) could not be evaluated in the binary predictor study.

Factors explaining the relative bias of unadjusted standard error estimates for the binary predictor are displayed in Table 4.10.

Table 4.10. *ANOVA for Relative Biases in SE_0 of γ_{02} for Binary Predictor Study*

Effect	Sum Sqr	Df	Mean Sqr	*F*	*p*	Partial η^2
J	4.050	2	2.025	334.346	<.001	.121
Binary predictor ratio	.103	1	.103	16.975	<.001	.003
Binary predictor effect	<.001	1	<.001	<.001	1.00	<.001
J x Binary predictor ratio	.044	2	.022	3.611	.027	.001
J x Binary predictor effect	<.001	2	<.001	<.001	1.00	<.001
Binary predictor ratio x Binary predictor effect	<.001	1	<.001	<.001	1.00	<.001
J x Binary predictor ratio x Binary predictor effect	<.001	2	<.001	<.001	1.00	<.001

Note. ANOVA = analysis of variance; SE_0 = unadjusted standard error estimates; *J* = number of clusters in sample.

Number of clusters, binary predictor ratio, and their interaction were the factors most strongly associated with the variability in relative bas of SE_0s estimates for the binary predictor effect (γ_{02}). As shown in Table 4.10, the difference in relative bias in the unadjusted *SE*s for γ_{02} between levels of the binary predictor effect (i.e., γ_{02} = .45 vs. γ_{02} = .20) was negligible. Consequently, relative bias, MSE, and RMSE from the binary predictor study were aggregated across levels of the binary predictor's effect.

Continuous Predictors' Effects

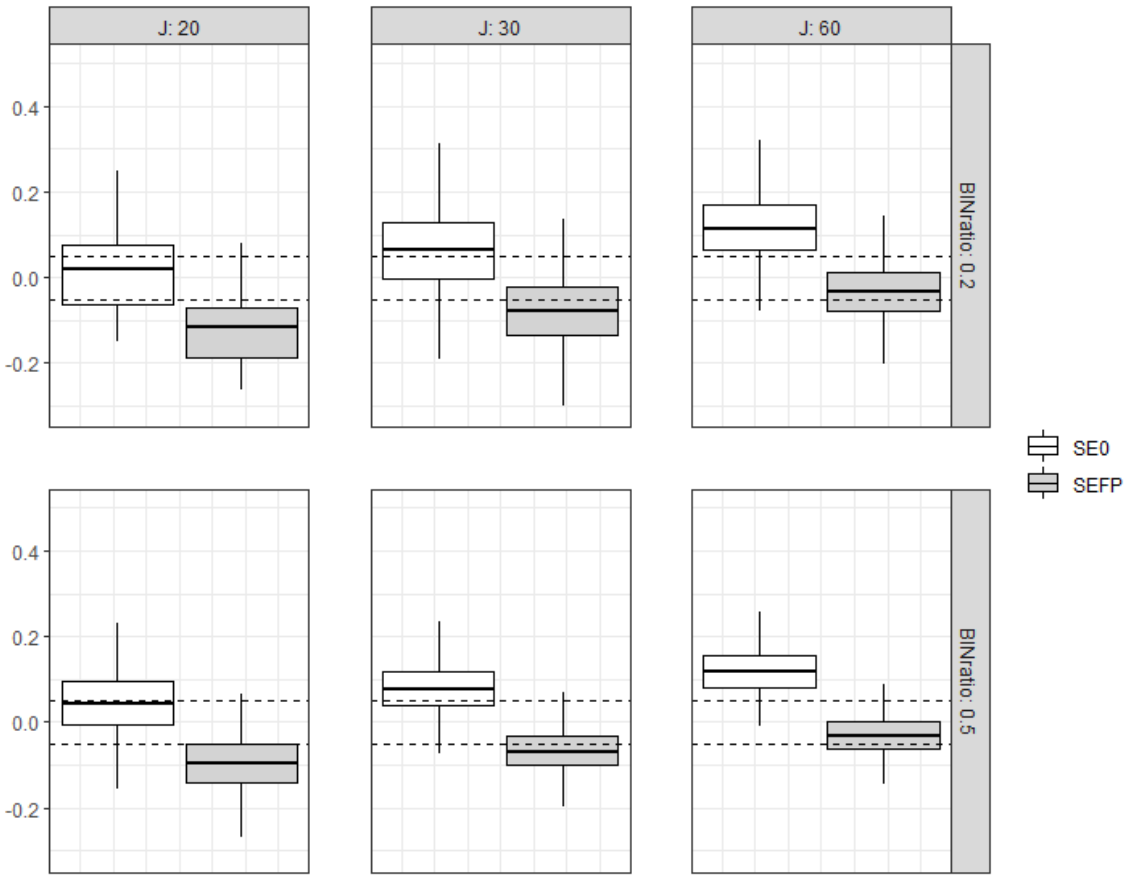
Estimates for the level-2 continuous predictor's effect (γ_{01}) and the level-1 continuous predictor's effect (γ_{10}) from the binary predictor study coincided with those from the continuous predictors study and are presented in Appendix C. The effect for the binary predictor is discussed in the following section.

Binary Predictor's Effect

γ_{02}. Figure 4.4 shows the percentage relative bias for γ_{02} in the binary predictor study. As shown below, the degree of relative bias in the unadjusted standard error estimates increased with larger *J*. The degree of relative bias in the unadjusted standard

error estimates was less when the binary predictor ratio was unbalanced (i.e., binary predictor ratio = .2) than when it was balanced (i.e., binary predictor ratio =.5), except when $J = 60$. However, the plots showing $J = 20$ and $J = 30$ when binary predictor ratio = .2 are based on fewer populations because non-convergence was an issue when drawing a small number of clusters from a population with an unbalanced binary predictor (see Table 4.8).

Figure 4.4. Percentage relative bias in *SE*s for γ_{02} in the binary predictor study.



The average relative biases for each condition shown in Figure 4.4 are also presented in Table 4.11. The FPC adjusted standard error estimates were negatively biased when $J = 20$. The SE^{FP}s were less biased than the unadjusted standard errors for a

balanced binary predictor ratio when $J \geq 30$. For the unbalanced binary predictor ratio, the SE^{FP}s were only acceptable when $J = 60$.

Table 4.11. *Average Relative Bias in SEs for γ_{02} in the Binary Predictor Study*

Binary predictor ratio	J	SE_0	SE^{FP}
.2	20	.026	-.111
	30	.068	-.075
	60	.119	-.031
.5	20	.045	-.095
	30	.081	-.064
	60	.121	-.029

Note. J = number of clusters in sample; SE_0 = unadjusted standard error estimates; SE^{FP} = FPC adjustment estimates.

MSE and RMSE for γ_{02} are presented in Table 4.12. As shown, the SE^{FP} estimates were closer to the empirical *SE*s than the SE_0 estimates across all conditions in the binary predictor study.

Table 4.12. *Mean Square Error and Root Mean Square Error for γ_{02} in Binary Predictor Study*

	MSE (RMSE)	
J	SE_0	SE^{FP}
20	.0244 (.1527)	.0233 (.1487)
30	.0154 (.1193)	.0127 (.1076)
60	.0053 (.0700)	.0028 (.0508)

Note. MSE = mean square error; RMSE = root mean square error; J = number of clusters in sample; SE_0 = unadjusted standard error estimates; SE^{FP} = FPC adjustment estimates. Values were rounded to the nearest .0001.

The binary predictor's effect only was also evaluated in terms of $Coverage_{\boldsymbol{\gamma}}$ and $Coverage_0$. Coverage is shown Table 4.13. As shown, interval estimates were more likely to include 0 when γ_{02} was small (i.e., $\gamma_{02} = .2$) than when it was medium (i.e., $\gamma_{02} = .45$). $Coverage_{\boldsymbol{\gamma}}$ rates for a small effect were identical to rates for a medium effect. Using SE^{FP} resulted in narrower interval estimates and fewer populations containing $\boldsymbol{\gamma}$ and 0.

Table 4.13. *Proportion of Populations with Interval Estimates Including* $\boldsymbol{\gamma}$ *and 0*

			$Coverage_{\boldsymbol{\gamma}}$		$Coverage_0$	
J	Binary predictor ratio	Binary predictor effect	SE_0	SE^{FP}	SE_0	SE^{FP}
20	.2	.2	.914	.867	.904	.857
		.45	.914	.867	.869	.813
	.5	.2	.912	.866	.901	.852
		.45	.912	.866	.857	.797
30	.2	.2	.925	.881	.911	.862
		.45	.925	.881	.860	.799
	.5	.2	.925	.881	.909	.859
		.45	.925	.881	.841	.775
60	.2	.2	.941	.900	.917	.868
		.45	.941	.900	.825	.750
	.5	.2	.937	.895	.908	.856
		.45	.937	.895	.780	.698

Note. J = number of clusters in sample; SE_0 = unadjusted standard error estimates; SE^{FP} = FPC adjustment estimates.

CHAPTER V. DISCUSSION

The purpose of the continuous predictors study was to evaluate the FPC adjustment in two-level hierarchical linear models for a few number of large clusters and compare the FPC adjustment's performance to a finite population bootstrapping alternative. The purpose of the binary predictor study was to examine the efficiency of the FPC adjustment for standard errors associated with a binary level-2 predictor. The following research questions were considered:

RQ1a. How do unadjusted standard errors (SE_0s) compare to FPC adjusted standard errors (SE^{FP}s)?

RQ1b. How do SE_0s compare to SE^{FP}s for data with a few number of large clusters (i.e., $J < 30$ & $n_j > 25$)?

RQ2. How do finite population bootstrapped standard errors (SE^{FPboot}s) compare to SE^{FP}s?

RQ3. How do SE_0s compare to SE^{FP}s for binary predictors?

RQ4. How do SE^{FPboot}s compare to SE^{FP}s binary predictors?

A Monte Carlo simulation was conducted to evaluate SE_0, SE^{FP}, and SE^{FPboot} using relative bias, mean squared error (MSE), and root mean square error (RMSE) as the dependent variables. A discussion of the current study's main findings, limitations and future directions, and implications for applied research is provided below.

Main Findings

Finite population corrections are seldom utilized in single-level studies because target populations in those studies tend to be so large that the sampling fraction is negligible. However, this is unlikely to be the case when generalizing findings to level-2

units, as in the examples of multilevel studies discussed in Chapter Two. Failing to account for a finite level-2 population in those studies may have resulted in erroneous interpretations of their results given the findings of the current study discussed throughout this chapter. The specific research questions are discussed in turn in the following sections.

In general, results of the current study indicated the *SE*s in two-level HLMs were positively biased when the assumption of infinite population was violated. Incorporating an additional design feature (i.e., the sampling fraction) into the model through the use of the FPC adjustment capitalized on the strengths of the integrated framework and alleviated much of the bias, resulting in smaller *SE*s than when the sampling fraction was ignored.

Applying FPC to HLMs and expressing the *SE*s of regression coefficients in terms of the finite population adjusted variance components is only one way to account for finite populations (Lai et al., 2018). Bootstrapping methods may prove to be viable alternatives to the FPC adjustment. However, the finite population bootstrap method utilized in the current study severely underestimated the empirical *SE*s.

RQ1a & b (Comparing SE_0s to SE^{FP}s for a Few Number of Large Clusters)

Lai et al. (2018) demonstrate how the degree of bias in unadjusted standard errors increases with larger cluster size. The largest cluster in their simulated conditions was 25. However, cluster sizes of at least 30 are normal in educational research (Mass & Hox, 2005; Maxwell et al., 2017; McNeish & Stapleton, 2016a; Oberle et al., 2011). The current study compared the performance unadjusted standard errors (SE_0s) to FPC

adjusted standard errors (SE^{FP}s) for data with large clusters (i.e., $n_j = 30,\ n_j = 90$, and $n_j = 150$).

Results showed SE_0s were overestimated as cluster size and number of clusters increased (see Figures 4.1 – 4.3). For the level-2 effects, the number of clusters (*J)* explained the most variability in unadjusted standard errors and the effect of cluster size (n_j) on SE_0s was negligible (see Tables 4.2 and 4.4). The FPC adjusted standard error estimates for the level-2 effects were less biased than the SE_0s when $J \geq 30$. This specific finding is important for applied researchers, especially those in educational settings, because students may be nested within a finite set of level-2 units greater than 30 (see Montague et al., 2014; Thrash & Warner, 2016).

There was substantial bias in the SE_0s for the level-1 effect across all samples size conditions. For the level-1 effect, the effect of cluster size (n_j) and its interaction with the number of clusters (J x n_j) explained more variability in unadjusted standard errors than J alone (see Table 4.6). The FPC correction removed virtually all the bias for the level-1 effect regardless of sample size (see Table 4.1 and Figure 4.3).

To summarize, the FPC adjusted standard error estimates exhibited acceptable levels of relative bias with large J and large n_j. SE^{FP}s performed better than SE_0s regardless of cluster size (n_j). This finding provided evidence that the FPC adjustment produces acceptable *SE*s for clusters larger than those simulated in Lai et al., (2018). Furthermore, SE^{FP}s performed better than SE_0s for large clusters when the number of clusters (J) was at least 30. SE^{FP}s did not perform well (i.e., were more biased than SE_0s) and were negatively biased for data with a few number of clusters ($J = 20$).

The main findings of RQ1a and b provide evidence that the FPC adjusted standard error estimates produce acceptable standard errors for cluster sizes common in educational settings, as long as the number of clusters is at least 30. This finding corresponds with suggestions for sufficient sample sizes for HLMs in general (e.g., Kreft and De Leeuw (1998) suggest 30 is the fewest acceptable number of clusters to obtain sufficient power in HLM).

RQ2 (Comparing SE^{FP}s to SE^{FPboot}s)

Lai et al. (2018) suggest comparing SE^{FP}s to bootstrapping alternatives. The finite population bootstrap estimates in the current study (SE^{FPboot}s) were obtained by drawing J' level-2 clusters (see Equation 2.31), and severely underestimated the empirical *SE*s across all conditions. The degree of relative bias in SE^{FPboot}s increased with larger J and larger n_j. The SE^{FPboot}s did not perform well in any of the conditions (i.e., the SE^{FPboot}s were more biased than SE_0s and SE^{FP}s; see Figures 4.1 – 4.3). The finite population bootstrap estimates for the level-1 effect (i.e., γ_{10}) were even more biased than those for the level-2 effects (see Table 4.1). Consequently, use of the finite population bootstrap is not suggested based on the main findings of RQ2.

RQ3 (Comparing SE_0s to SE^{FP}s for Binary Predictor)

Lai et al. (2018) did not compare SE_0s to SE^{FP}s for binary predictors. Findings from the current study suggest standard errors for a relatively balanced binary level-2 predictor function similarly in terms of bias as continuous predictors.

The number of clusters (*J*) explained the most variability in unadjusted standard errors for the binary predictor. Relative bias in the SE_0s increased with larger *J* (see

Figure 4.4). The binary predictor ratio and its interaction with *J* explained a substantial proportion of variability in unadjusted standard errors (see Table 4.10).

The average relative bias in the SE_0s for a balanced binary predictor (i.e., binary predictor ratio = .5) was larger than that for an unbalanced binary predictor (i.e., binary predictor ratio = .2) when $J = 20$ or 30 (see Table 4.11). However, those estimates of the average relative bias in SE_0s for an unbalanced binary predictor were based upon conditions with a few number of populations with complete sample replications (see Table 4.8). The average relative bias in SE_0s for a balanced binary predictor was smaller than that for an unbalanced binary predictor when $J = 60$.

The standard errors for a relatively balanced binary predictor ratio functioned similarly in terms of bias as the continuous predictors. The average relative bias of the FPC adjusted standard errors for a balanced binary predictor were less than the relative bias in SE_0s when $J \geq 30$. Similar to Bell et al., (2013) as cited in McNeish & Stapleton (2016b), unbalanced binary predictors required larger *J* to achieve acceptable standard error estimates. The FPC adjusted standard errors for an unbalanced binary predictor exhibited more bias than unadjusted *SE*s when $J < 60$ (see Figure 4.4 and Table 4.11).

The main findings of RQ3 corroborate those from McNeish and Stapleton (2016a), and McNeish and Stapleton (2016b) and provide evidence that *SE*s for a balanced binary level-2 predictor function similarly as continuous predictors, whereas unbalanced binary level-2 predictors require a larger number of clusters to achieve acceptable *SE*s. Although not explicitly tested in the current simulation, the FPC adjustment is expected to produce acceptable *SE*s for balanced binary level-1 predictors because findings from the current study suggested *SE*s for a balanced binary predictor

function similarly as continuous predictors and the FPC adjustment removed virtually all the bias for a *continuous* level-1 predictor (see Table 4.1 and Figure 4.3). The FPC adjustment may yield acceptable estimates of standard errors for an unbalanced binary level-1 predictor, but likely requires a greater number of clusters to obtain acceptable estimates than needed for a balanced binary predictor.

RQ4 (Comparing SE^{FP}s to SE^{FPboot}s for Binary Predictor)

The finite population bootstrap estimates (SE^{FPboot}s) could not be evaluated in the binary predictor study because non-convergence in the bootstrapped samples was almost certain as the number of clusters decreased and when the binary predictor was unbalanced (see Table 4.9). Consequently, SE^{FPboot}s could not be compared to SE^{FP}s for binary predictors.

Had the models for the bootstrapped samples been able to convergence, SE^{FPboot}s were not expected to produce acceptable standard errors for a binary predictor because they severely underestimated the empirical *SE*s (see Figures 4.1 – 4.3) and were more biased than SE_0s in the continuous predictors study (see Table 4.1). However, based solely on the evidence from current study, it remains unclear how SE^{FPboot}s compare to SE^{FP}s binary predictors.

Limitations & Future Directions

The main findings of the current study must be considered in light of its limitations. First, the finite population bootstrapping procedure chosen for the current study resamples J' level-2 units. However, the procedure used was not designed for HLM (Davison & Hinkley, 1997) and may have resulted in larger sample sizes than its single-level alternative (i.e., resampling n' level-1 units only). Because J' is greater than J (see

Equation 2.31) resampling J' level-2 units likely results in larger bootstrapped samples and smaller standard error estimates than what would have been obtained using J level-2 units. Future study is needed to determine if *SE*s obtained from resampling J level-2 units are more appropriate than those obtained from resampling J' level-2 units from finite populations.

Second, the finite population bootstrapping procedure chosen for the current study sampled only complete level-2 units. Although SE^{FPboot}s underestimated the empirical standard errors across conditions, their average relative bias was more severe for the level-1 effect (i.e., γ_{10}) than for the level-2 effects (i.e., γ_{01} and γ_{02}). See Table 4.1. Future study is needed to determine if drawing n_j level-1 units with replacement from each level-2 unit alleviates any relative bias associated with level-1 effects.

Third, the finite population bootstrapping procedure chosen for the current study is a nonparametric bootstrap method because it samples from the empirical distribution of the observed data (Davison & Hinkley, 1997; Efron, 1979). Use of the nonparametric, finite population bootstrapping procedure chosen did not result in acceptable *SE*s (i.e., average relative bias in SE^{FPboot}s exceeded average relative bias in SE_0s across all simulated conditions). Future study is needed to compare the FPC adjustment to other resampling methods, such as parametric and parametric residual bootstrapping procedures.

Fourth, only populations with complete sample replications were included in the binary predictor study (see Table 4.8). Future research may benefit from drawing additional samples to ensure accurate comparisons across samples with different numbers of clusters (J). However, doing so would increase the computational load of generating

and analyzing data. Furthermore, additional samples drawn would not be selected truly at random, adding complexity to the sampling design implemented.

Fifth, the current study assumed there were no nonsampling errors (e.g., no measurement error). Future study is needed to evaluate the FPC adjustment in HLMs with latent variables. Nonsampling errors also encompass situations in which the actual probabilities of selection differ from those of the presumed design (Thompson, 2012). The current study assumed probabilities of inclusion were known and equal for each sample drawn, and future study is needed to evaluate the FPC adjustment when samples possess unequal probabilities of selection.

Sixth, conclusions drawn are limited to the conditions included in the simulation. The current binary predictor study held sample-population ratio, cluster size, and ICC constant. Future study is needed to determine how *SE*s for binary predictors interact with sample-population ratio, cluster size, and ICC. Furthermore, the current study only examined a binary level-2 predictor. Evidence suggests the FPC adjustment produces acceptable *SE*s for a binary level-2 predictor, although incorporating an unbalanced predictor into HLMs may require greater number of clusters. Future study is needed to determine if the FPC adjustment produces acceptable *SE*s for binary level-1 predictors (e.g., gonosome and English language learner designation). Additionally, the current study does not include any cross-level interactions. Examination of *SE*s for cross level-interactions with continuous and binary predictors in finite populations also warrants future study.

Finally, the current study demonstrates how to use the FPC adjustment to correct *SE*s in two-level models. However, more complex data structures are likely to be

encountered in applied settings. Misspecification of a model for a population with cross-classified or multiple membership structure can lead to inaccurate estimates of standard errors (Lou & Kwok, 2009). Future study is need to extend and evaluate the FPC adjustment for data structures with more than two levels, with cross-classification, and with multiple group memberships.

Implications for Applied Research

The current study evaluated the FPC adjustment in two-level hierarchical linear models for a few number of large clusters, compared the FPC adjustment's performance to a finite population bootstrapping alternative, and examined the efficiency the FPC adjustment for standard errors associated with a binary level-2 predictor. Although based on simulated data, the findings offer several important implications for applied research.

The FPC adjusted standard error estimates for the level-2 effects exhibited acceptable levels of relative bias across most conditions. Average relative bias in SE^{FP}s is less than the relative bias in SE_0s regardless of cluster size when $J \geq 30$. Relative bias in SE^{FP}s was less than 5% when $J = 60$, and was less than the relative bias in SE_0s when $J = 30$. However, SE^{FP}s underestimated the empirical *SE*s and were more biased than SE_0s when $J = 20$ (see Figures 4.1 and 4.2). Consequently, it is suggested that the FPC adjustment only be applied when then number of clusters is at least 30. This suggestion corresponds with Kreft and De Leeuw's (1998) guideline of 30 clusters to obtain sufficient power in HLM. Based on this guideline, if applied researchers have sufficient samples sizes for HLM, they also have sufficient samples sizes to consider the FPC adjustment.

There was substantial bias in SE_0s for the level-1 effect (i.e., γ_{10}) across conditions in the continuous predictors study. Applying the FPC correction at both levels removed virtually all that bias (see Figure 4.3). In light of this finding, applied researchers are encouraged to correct for finite level-2 units even if they are only interested in level-1 predictors.

The finite population bootstrap procedure underestimated the empirical *SE*s across conditions. Ergo, the finite population bootstrap procedure is not recommended for use in applied settings. Rather, applied researchers should rely upon the FPC adjustment when needed.

*SE*s for a balanced binary level-2 predictor (i.e., binary predictor ratio = .5) functioned similarly in terms of bias as continuous predictors. The relative bias in SE^{FP}s for a balanced predictor was smaller than the relative bias in SE_0s when $J = 30$ or $J = 60$. However, SE^{FP}s for a balanced binary predictor were more biased than SE_0s when $J = 20$. Consequently, it is recommended that the FPC adjustment be applied to level-2 balanced binary predictors' effects when the number of clusters is at least 30, which corresponds with Kreft and De Leeuw's (1998) sufficient sample size guidelines for HLM.

More clusters are needed when estimating standard errors for an unbalanced binary predictor (i.e., binary predictor ratio = .2). Standard errors associated with an unbalanced binary predictor exhibited bias when based on fewer than 60 clusters. For an unbalanced binary predictor, SE^{FP}s were only acceptable when $J = 60$. Therefore, it is recommended that the FPC adjustment only be applied to unbalanced binary predictors' effects when the number of clusters is at least 60.

Applied researchers, particularly in education, should consider incorporating FPCs into their HLMs because higher level units common in educational settings (e.g., schools, classrooms, and districts) are likely to be considered finite, as in the examples discussed in Chapter Two. Applied research questions related to those higher level units may involve binary predictors (e.g., public vs. private schools). Main findings from the current study provide evidence that the FPC adjustment produces acceptable standard error estimates for balanced binary predictors as well as continuous predictors. However, FPCs for HLMs including unbalanced binary predictors may require a greater number of clusters.

In conclusion, results from the current study indicate standard error estimates for continuous and binary predictors in HLMs are positively biased when a finite population is ignored. The relative bias in SE_0s increases with greater numbers of clusters and larger cluster sizes. Use of the FPC adjustment generally results in smaller *SE*s. Consequently, applied researchers may abuse the FPC adjustment by arbitrarily redefining a given population to deflate the uncertainty of their estimates and to obtain statistical significance (Lai et al., 2018). To combat such practice, applied researchers need to choose and define their target population *a priori*. Every practical statistician must ask "of what population is this a random sample?" (Fisher, 1922, p. 313). Applied researchers should identify and explicitly state their target populations, examine their sampling fraction, and consider using the FPC adjustment because doing so yields more accurate inferences for finite populations.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1-48.

Chernick, M. R. (1999). *Bootstrap methods a practitioner's guide*. New York, NY: Wiley.

Chernick, M. R., & LaBudde, R. A. (2011). *An introduction to bootstrap methods with applications to R*. Hoboken, NJ: Wiley.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York, NY: Wiley.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge, UK: Cambridge University Press.

Dedrick, R. F., Ferron, J. M., Hess, M. R. Hogarty, K. Y., Kromrey, J. D., Lang, T. R., Niles, J. D., & Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research, 79*, 69-102. doi: 10.3102/0034654308325581

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*, 1-26.

Efron, B. (1982, June). *The jackknife, the bootstrap, and other resampling plans*. Presented at the CBMS-NSF Regional Conference Series in Applied Mathematics. SAIM, Philadelphia.

Efron, B. (1993). *An introduction to the bootstrap.* New York, NY: Chapman & Hall.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician, 37*, 36-48.

Fisher, R. A. (1922). On the mathematical foundation of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A, 222*, 309-368.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society, B, 17*, 69-78.

Fisher, R. A. (1956). *Statistical methods and scientific inference.* New York, NY: Hafner.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multi-level/hierarchical models.* New York, NY: Cambridge University Press.

Goldstein, H. (2011). Bootstrapping in multilevel models. In J.J. Hox, & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 163-171). New York, NY: Routledge.

Green, B. F. Jr., & Tukey, J. W. (1960). Complex analyses of variance: General problems. *Psychometrika, 25*, 127-152.

Hansen, M., Madow, W., & Tepping, B. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association, 78*, 776-793.

Hofert, M., Kojadinovic, I., Maechler, M., & Yan, J. (2018). *copula: Multivariate dependence with copulas*. R package version 0.999-19. https://CRAN.R-project.org/package=copula.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and meta-analysis. *Sociological Methods & Research, 26*, 329-367.

How many countries are there in the world? (2018). Retrieved from https://www.worldatlas.com/articles/how-many-countries-are-in-the-world.html

Izrailev, S. (2014). *tictoc: Functions for timing R scripts, as well as implementations of stack and list structures*. R package version 1.0. https://CRAN.R-project.org/package=tictoc.

Johnstone, D. J. (1987). Tests of significance following R. A. Fisher. *British Journal of Philosophy of Science, 38*, 481-499.

Juvonen, J., Wang, Y., & Espinoza, G. (2011). Bullying experiences and compromised academic performance across middle school grades. *Journal of Early Adolescence, 31*, 152-173. doi: 10.1177/0272431610379415

Kish, L. (1965). *Survey sampling*. New York, NY: Wiley.

Kish, L. (2003). The hundred years' wars of survey sampling. In Kalton, G. & Heeringa, S. (Eds.), *Leslie Kish Selected Papers* (pp. 5-19). Hoboken, NJ: Wiley. (Reprinted from *Statistics in Transition, 2*, pp. 813-830, by L. Kish, 1995)

Kojadinovic, I., & Yan, J. (2010). Modeling multivariate distributions with continuous margins using the copula R package. *Journal of Statistical Software, 34*, 1-20.

Kreft, I., & De Leeuw, J. (1998). *Introducing multilevel modeling*. London, England: Sage.

Lai, M. H. C., Kwok, O., Hsiao, Y., & Cao, Q. (2018). Finite population correction for two-level hierarchical linear models. *Psychological Methods, 23*, 94-112.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association, 88*, 1242-1249.

Little, R. J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association, 99*, 546-556. doi: 10.1198/016214504000000467

Lou, W., & Kwok, O. (2009). The impacts of ignoring a crossed factor in analyzing cross-classified data. *Multivariate Behavioral Research, 44*, 182-212. doi: 10.1080/00273170902794214

Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica, 58*, 127-137.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*, 86-92. doi: 10.1027/1614-1881.1.3.86

Mani, S., Anita, K. D., & Rindfleisch, A. (2007). Entry mode and equity level: A multilevel examination of foreign direct investment ownership structure. *Strategic Management Journal, 28*, 857-866. doi: 10.1002/smj.611

Maxwell, S., Reynolds, K. J., Lee, E., Subasic, E., & Bromhead, D. (2017). The impact of school climate and school identification on academic achievement: Multilevel Modeling with student and teacher data. *Frontiers in Psychology, 8*, 1-21. doi: 10.3389/fpsyg.2017.02069

McNeish, D. (2017). Small sample methods for multilevel modeling: A colloquial elucidation of REML and the Kenward-Roger correction. *Multivariate Behavioral Research, 52*, 661-670. doi: 10.1080/00273171.2017.1344538

McNeish, D., & Stapleton, L. M. (2016a). Modeling clustered data with very few clusters. *Multivariate Behavioral Research, 51*, 495-518. doi: 10.1080/00273171.2016.1167008

McNeish, D. M., & Stapleton, L. M. (2016b). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review, 28*, 295-314. doi: 10.1007/s10648-014-9287-x

Montague, M., Krawec, J., Enders, C., & Dietz, S. (2014). The effects of cognitive strategy instruction on math problem solving of middle-school students of varying ability. *Journal of Educational Psychology, 2*, 469-481. doi: 10.1037/a0035176

Mostafa, M. (2013). Wealth, post-materialism and consumers' pro-environmental intentions: A multilevel analysis across 25 nations. *Sustainable Development, 21*, 385-399. doi: 10.1002/sd.517

Nelsen, R. B. (1999). *An introduction to copulas*. New York, NY: Springer.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the royal Statistical Society, 109*, 558-606.

Nielsen, S. (2009). Why do top management look the way they do? A multilevel exploration of the antecedents of TMT heterogeneity. *Strategic Organization, 7*, 277-305. doi: 10.1177/147612700934049

Oberle, E., Schonert-Reichl, K. A., & Zumbo, B. (2011). Life satisfaction in early adolescence: Personal, neighborhood, school, family, and peer influences. *Journal of Youth and Adolescence, 40*, 889-901. doi: 10.1007/s10964-010-9599-1

Peretz, H., & Fried, Y. (2012). National cultures, performance appraisal practices, and organizational absenteeism and turnover: A study across 21 countries. *Journal of Applied Psychology, 97*, 448-459. doi: 10.1037/a0026011

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology, 48*, 85-112. doi: 10.1016/j.jsp.2009.09.002

Peugh, J. L., & Enders, C. K. (2005). Using the SPSS MIXED procedure to fit cross-sectional and longitudinal multilevel models. *Educational and Psychological Measurement, 65*, 717-741. doi: 10.1177/0013164405278558

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review, 61*, 317-337.

Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research, 5*, 239-261.

R Core Team. (2019). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing: Vienna, Austria. Retrieved from http://www.R-project.org.

RStudio Team (2018). *RStudio: Integrated development for R.* RStudio, INC.: Boston, MA: Retrieved from http://www.rstudio.com.

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society, Series A, 169*, 805-827.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: SAGE.

Raykov, T. (2010). Proportion of third-level variation in multi-level studies: A note on an interval estimation procedure. *British Journal of Mathematical and Statistical Psychology, 63*, 417-426. doi: 10.1348/000711009X468004

Raykov, T., Patelis, T., Marcoulides, G. A., & Lee, C. (2016). Examining intermediate omitted levels in hierarchical designs via latent variable modeling. *Structural Equation Modeling, 23*, 111-115. doi: 10.1080/10705511.2014.938186

Rockstuhl, T., Dulebohn, J. H., Ang, S., & Shore, L. M. (2012). Leader-member exchange (LMX) and culture: A meta-analysis of correlates of LMX across 23 countries. *Journal of Applied Psychology, 6*, 1097-1130. doi: 10.1037/a0029978

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 593-604.

Särndal, C. –E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York, NY: Springer.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance Components*. New York: Wiley.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Smith, T. M. F. (1994). Sample surveys 1975-1990: An age of reconciliation? *International Statistical Review/Revue Internationale de Statistique, 65*, 5-19.

Snijders, T. A. B. (2005). Power and sample size in multilevel linear models. In B. S. Everitt, & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1570-1573). Chichester, UK: Wiley.

Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics, 18*, 237-59.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Thousand Oaks, CA: SAGE.

Sterba, S. K. (2009). Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration. *Multivariate Behavioral Research, 44*, 711-740. doi: 10.1080/00273170903333574

Sugden, R. A., & Smith, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika, 71*, 495-506.

Swierzy, P., Wicker, P., & Breuer, C. (2018). Usefulness of multilevel modeling in sport management research: The case of voluntary roles in nonprofit sports clubs. *Measurement in Physical Education and Exercise Science, 22*, 1-12. doi: 10.1080/1091367X.2018.1438289

Thrash, C. R., & Warner, T. D. (2016). The geography of normative climates: An application to adolescent substance use. *Journal of Youth and Adolescence, 45*, 1587-1603. doi: 10.1007/s10964-016-0444-z

Thomas, S. L., & Heck, R. H. (2001). Analysis of large-scale secondary data in higher education research: Potential perils associated with complex sampling designs. *Research in Higher Education, 42*, 517-540.

Thompson, S. K. (2012). *Sampling* (3rd ed.). Hoboken, NJ: Wiley.

van der Leeden, R., Meijer, E., & Busing, F. M. T. A. (2007). Resampling multilevel models. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 403-435). New York, NY: Springer.

Yan, J. (2007). Enjoy the joy of copulas: With a package copula. *Journal of Statistical Software, 21*, 1-21.

APPENDIX A: ABBREVIATIONS & NOTATION

SRS	-	Simple random sampling
HLM	-	Hierarchical linear modeling
FPCs	-	Finite population corrections
PSUs	-	Primary sampling units
iid	-	Identically and independently distributed random variables
ICC	-	Intraclass correlation coefficient
*SE*s	-	Standard errors
π_is	-	Probability of selection
f	-	Sampling fraction
P	-	Sample to population ratio
J	-	Number of clusters in sample
J_{pop}	-	Number of clusters in finite population
n_j	-	Number of elements drawn from each cluster
N_j	-	Number of elements in finite population
SE_0s	-	Unadjusted standard errors
SE^{FP}s	-	FPC adjusted standard errors
SE^{FPboot}s	-	Finite population bootstrapped standard errors
W_{2j}^{B}	-	Binary level-2 predictor
F_n	-	Empirical distribution of a parameter F
$F_n{}^*$	-	Bootstrap distribution sampling with replacement from F_n

APPENDIX B: R CODE USED FOR SIMULATION

```
##########LOAD/INSTALL PACKAGES: tictoc, copula, & lme4##########

#install.packages('tictoc')
#install.packages('copula')
#install.packages('lme4')
library(tictoc)
library(copula)
library(lme4)

#tic("total")#start timer

##########SETTING UP SIMULATION CONDITIONS##########

pop_id <- commandArgs(trailingOnly=TRUE)[1]

set.seed(3916 + as.numeric(pop_id)) #Happy Birthday Hadley James!
#set.seed(62818) #Happy Birthday Huxley Timber!

POPreps<-1
SAMPreps<-500
BOOTreps<-500

#Setting up Sample Size Conditions:
P<-.25 #manipulate sample-to-population ratio (P)
J<-60 #manipulate number of clusters in sample
Jpop<-J/P #number of clusters in population
nj<-150 #manipulate cluster size

Jboot<-ceiling((J-1)/(1-P))#rounded up to nearest integer

#Population Parameters:

#Fixed effects From Lai et al. (2018):
gam00<-0 #intercept fixed at zero

#Remember: Transpose effects for binary predictor study!!!
gam01<-.2 #fixed L2small effect
gam02<-.45 #fixed L2 medium effect

#Setting up ICC=tau00/(tau00+sigmasquare)
tau00<-1 #i.e., variance of 'u0j'
ICC<-.2
#ICC equation rewritten for sigmasquare:
sigmasquare<-(tau00-(ICC*tau00))/ICC #i.e., variance of 'eij'
```

```
gam10<-.577*sqrt(sigmasquare)

tau11<-.5 #i.e., variance of 'u1j'

#setting up copula for desired covariance among L2 predictors-USE 'binary gen
test2.R' to experiment with parameters!
#MAKE SURE TO USE CORRECT COPULA FOR PREDICTOR TYPE!!!!

#myCop<-normalCopula(param = c(.625,0,0,0,0,.25), dim = 4, dispstr =
"un")#copula for binaray
myCop<-normalCopula(param = c(.5,0,0,0,0,.25), dim = 4, dispstr = "un")#copula for
continuous

#MAKE SURE TO Change binary predictor discrepancy here:
BINratio<-0 #0 for continuous label

##########START of  FPC Function adapted from Lai et al. (2018)##########
vcovFPC <- function(. = NULL, popsize2 = NULL,
                    popsize1 = NULL) {

 if (!inherits(., "merMod")) {
   stop("Wrong input: Not a fitted model from lmer() with class merMod")
 }
 if (length(.@flist) != 1) {
   stop("Wrong input: Only models with two levels are supported")

 }
 if (is.null(popsize1) & is.null(popsize2)) {
   message("No FPC specified; return results from lme4::vcov.merMod()")
   return(vcov(.))
 }
 PR <- .@pp
 N <- unname(.@devcomp$dims["n"])
 nclus <- unname(ngrps(.))
 if (isTRUE(popsize2 > nclus)) fpc2 <- 1 - nclus / popsize2
 else {
   fpc2 <- 1
   message("No FPC needed at Level-2")
 }
 if (isTRUE(popsize1 > N)) fpc1 <- 1 - N / popsize1
 else {
   fpc1 <- 1
   message("No FPC needed at Level-1")
 }
 if (fpc1 == 1 & fpc2 ==1) {
```

```
   message("Return results from lme4::vcov.merMod()")
   return(vcov(.))
  }
  A<- PR$Lambdat %*% PR$Zt
  Astar <- A * sqrt(fpc2)
  X <- PR$X
  Astar_X <- Astar %*% X
  D <- Matrix::Diagonal(nrow(Astar), fpc1) + tcrossprod(Astar)
  Fisher_I <- (crossprod(X) - crossprod(solve(t(chol(D)), Astar_X))) / fpc1
  Phi <- solve(Fisher_I) * sigma(.)^2
  Phi <- as(Phi, "dpoMatrix")
  nmsX <- colnames(X)
  dimnames(Phi) <- list(nmsX, nmsX)
  return(Phi)
}
##########END of FPC Function adapted from Lai et al. (2018)##########

#NEED TO SAVE CONDITION RESULTS (i.e., 1 row for each pop with averages from
samples)
CONDresults<-matrix(nrow=POPreps, ncol=39) ###REFORMAT LATER
colnames(CONDresults)<-c("popID", "binary", "Wj2effect", "BINratio","numcluster",
"clustersize", "rb0gam01", "rb0gam02", "rb0gam10", "rbFPgam01", "rbFPgam02",
"rbFPgam10","rbBOOTgam01", "rbBOOTgam02", "rbBOOTgam10", "mse0gam01",
"mse0gam02", "mse0gam10", "mseFPgam01", "mseFPgam02",
"mseFPgam10","mseBOOTgam01", "mseBOOTgam02", "mseBOOTgam10",
"Rmse0gam01", "Rmse0gam02", "Rmse0gam10", "RmseFPgam01",
"RmseFPgam02", "RmseFPgam10","RmseBOOTgam01", "RmseBOOTgam02",
"RmseBOOTgam10","ZEROcover0gam02", "ZEROcoverFPgam02",
"ZEROcoverBOOTgam02", "PARAMcover0gam02", "PARAMcoverFPgam02",
"PARAMcoverBOOTgam02")

##########START of POPULATION GENERATION CODE##########


counterPOP<-0

#population level:
for (r in 1:POPreps){ #start POPreps loop
  counterL2<-0
  counterL1<-0
  counterPOP<-counterPOP+1

  POP.data<-matrix(nrow=Jpop*nj, ncol=14)
  colnames(POP.data)<-c("popID", "L2ID", "L1ID", "gam00", "gam01", "W1j",
"gam02", "W2j", "gam10","Xij", "u1j", "u0j", "eij", "yij" )
```

```
 #LEVEL-2:
 for (j in 1:Jpop){ #start L2 loop
  counterL2<-counterL2+1

  #using copula to generate predictors
  tempdata<-rCopula(1, myCop)
  colnames(tempdata)<-c("W1j", "W2j", "u0j", "u1j")
  W1j<-qnorm(tempdata[,"W1j"], mean=0, sd=sqrt(1))

  #MAKE SURE W2j matches correct copula above: comment out the other!!!
  #W2j<-qbinom(tempdata[,"W2j"], size = 1, prob = BINdescrepancy)#binary
  W2j<-qnorm(tempdata[,"W2j"], mean = 0, sd=(sqrt(1)))#continuous

  u0j<-qnorm(tempdata[,"u0j"], mean=0, sd=sqrt(tau00))
  u1j<-qnorm(tempdata[,"u1j"], mean=0, sd=sqrt(tau11))

  #LEVEL-1:
  for (n in 1:nj){ #start L1 loop
   counterL1<-counterL1+1

   Xij<-rnorm(n=1, 2, sqrt(1))
   eij<-rnorm(n=1, 0, sqrt(sigmasquare))
   yij<-gam00+gam01*W1j+gam02*W2j+gam10*Xij+u1j*Xij+u0j+eij

   POP.data[counterL1, "popID"]<-counterPOP
   POP.data[counterL1, "L2ID"]<-counterL2
   POP.data[counterL1, "L1ID"]<-counterL1
   POP.data[counterL1, "gam00"]<-gam00
   POP.data[counterL1, "gam01"]<-gam01
   POP.data[counterL1, "W1j"]<-W1j
   POP.data[counterL1, "gam02"]<-gam02
   POP.data[counterL1, "W2j"]<-W2j
   POP.data[counterL1, "gam10"]<-gam10
   POP.data[counterL1, "Xij"]<-Xij
   POP.data[counterL1, "u1j"]<-u1j
   POP.data[counterL1, "u0j"]<-u0j
   POP.data[counterL1, "eij"]<-eij
   POP.data[counterL1, "yij"]<-yij
  }# L1 loop END
 }# l2 loop END

 #SAVING POPULATION DATA
 Jpopdata<-as.data.frame(POP.data)

##########START of SAMPLING CODE##########
```

```
 #creates matrix for results from SRS (i.e., without replacement)
 SAMPresults<-matrix(nrow=SAMPreps, ncol=19)
 colnames(SAMPresults)<-c("sampID", "gam01", "gam02", "gam10", "se0gam01",
"se0gam02", "se0gam10", "seFPgam01", "seFPgam02", "seFPgam10",
"seBOOTgam01", "seBOOTgam02", "seBOOTgam10", "ZEROcover0gam02",
"ZEROcoverFPgam02", "ZEROcoverBOOTgam02", "PARAMcover0gam02",
"PARAMcoverFPgam02", "PARAMcoverBOOTgam02")

 counterSAMP<-0

 for (s in 1:SAMPreps){

  counterSAMP<-counterSAMP+1

  tempsamp<-sample(unique(Jpopdata$L2ID),J, replace = FALSE)
  sampdata<-subset(Jpopdata, L2ID %in% tempsamp)

  samplemodel<-lmer(yij ~ 1 + W1j + W2j + Xij + (1+Xij|L2ID), data = sampdata,
REML=TRUE)

  FEmat<-fixef(samplemodel)
  SE0mat<-sqrt(diag(vcov(samplemodel)))
  SEFPmat<-sqrt(diag(vcovFPC(samplemodel, Jpop, Jpop*nj)))

  #saving results from each sample
  SAMPresults[counterSAMP, "sampID"]<-counterSAMP
  SAMPresults[counterSAMP, "gam01"]<-FEmat[2]
  SAMPresults[counterSAMP, "gam02"]<-FEmat[3]
  SAMPresults[counterSAMP, "gam10"]<-FEmat[4]
  SAMPresults[counterSAMP, "se0gam01"]<-SE0mat[2]
  SAMPresults[counterSAMP, "se0gam02"]<-SE0mat[3]
  SAMPresults[counterSAMP, "se0gam10"]<-SE0mat[4]
  SAMPresults[counterSAMP, "seFPgam01"]<-SEFPmat[2]
  SAMPresults[counterSAMP, "seFPgam02"]<-SEFPmat[3]
  SAMPresults[counterSAMP, "seFPgam10"]<-SEFPmat[4]

  #creates matrix for results from FPbootstrap (i.e., with replacement)
  BOOTresults<-matrix(nrow=BOOTreps, ncol=4)
  colnames(BOOTresults)<-c("bootID","seBOOTgam01", "seBOOTgam02",
"seBOOTgam10")

  counterBOOT<-0

   for (b in 1:BOOTreps){

    counterBOOT<-counterBOOT+1
```

```
      cls <- sample(unique(sampdata$L2ID), Jboot, replace=TRUE)
      cls.col <- data.frame(L2ID=cls)

      # reconstructing the sample
      bootdata<-merge(cls.col, sampdata, by="L2ID")

      bootmodel<-lmer(yij ~ 1 + W1j + W2j + Xij + (1+Xij|L2ID), data = bootdata,
REML=TRUE)

      FEBOOTmat<-fixef(bootmodel)

      #saving results from each bootstrap replication
      BOOTresults[counterBOOT, "bootID"]<-counterBOOT
      BOOTresults[counterBOOT, "seBOOTgam01"]<-FEBOOTmat[2]
      BOOTresults[counterBOOT, "seBOOTgam02"]<-FEBOOTmat[3]
      BOOTresults[counterBOOT, "seBOOTgam10"]<-FEBOOTmat[4]

    }#BOOTreps loop END

   SAMPresults[counterSAMP, "seBOOTgam01"]<-
sd(BOOTresults[,"seBOOTgam01"])
   SAMPresults[counterSAMP, "seBOOTgam02"]<-
sd(BOOTresults[,"seBOOTgam02"])
   SAMPresults[counterSAMP, "seBOOTgam10"]<-
sd(BOOTresults[,"seBOOTgam10"])

   SAMPresults[counterSAMP, "ZEROcover0gam02"]<-
ifelse((SAMPresults[counterSAMP, "gam02"])-
1.96*(SAMPresults[counterSAMP,"se0gam02"])<0 &(SAMPresults[counterSAMP,
"gam02"])+1.96*(SAMPresults[counterSAMP,"se0gam02"])>0, yes = 1, no = 0)
   SAMPresults[counterSAMP, "ZEROcoverFPgam02"]<-
ifelse((SAMPresults[counterSAMP, "gam02"])-
1.96*(SAMPresults[counterSAMP,"seFPgam02"])<0 &(SAMPresults[counterSAMP,
"gam02"])+1.96*(SAMPresults[counterSAMP,"seFPgam02"])>0, yes = 1, no = 0)
   SAMPresults[counterSAMP, "ZEROcoverBOOTgam02"]<-
ifelse((SAMPresults[counterSAMP, "gam02"])-
1.96*(SAMPresults[counterSAMP,"seBOOTgam02"])<0
&(SAMPresults[counterSAMP,
"gam02"])+1.96*(SAMPresults[counterSAMP,"seBOOTgam02"])>0, yes = 1, no = 0)
   SAMPresults[counterSAMP, "PARAMcover0gam02"]<-
ifelse((SAMPresults[counterSAMP, "gam02"])-
1.96*(SAMPresults[counterSAMP,"se0gam02"])<gam02
&(SAMPresults[counterSAMP,
"gam02"])+1.96*(SAMPresults[counterSAMP,"se0gam02"])>gam02, yes = 1, no = 0)
```

```
   SAMPresults[counterSAMP, "PARAMcoverFPgam02"]<-
ifelse((SAMPresults[counterSAMP, "gam02"])-
1.96*(SAMPresults[counterSAMP,"seFPgam02"])<gam02
&(SAMPresults[counterSAMP,
"gam02"])+1.96*(SAMPresults[counterSAMP,"seFPgam02"])>gam02, yes = 1, no =
0)
   SAMPresults[counterSAMP, "PARAMcoverBOOTgam02"]<-
ifelse((SAMPresults[counterSAMP, "gam02"])-
1.96*(SAMPresults[counterSAMP,"seBOOTgam02"])<gam02
&(SAMPresults[counterSAMP,
"gam02"])+1.96*(SAMPresults[counterSAMP,"seBOOTgam02"])>gam02, yes = 1, no
= 0)
   }#SAMPreps loop END

  #saving results from each population
  #CONDresults[counterPOP, "popID"]<-pop_id
  CONDresults[counterPOP, "binary"]<-0 #REMEMBER TO CHANGE zero for
continuous, 1 for binary predcitor study!!!
  CONDresults[counterPOP, "Wj2effect"]<-gam02
  CONDresults[counterPOP, "BINratio"]<-BINratio
  CONDresults[counterPOP, "numcluster"]<-J
  CONDresults[counterPOP, "clustersize"]<-nj

  CONDresults[counterPOP, "rb0gam01"]<-(mean(SAMPresults[,"se0gam01"])-
sd(SAMPresults[,"gam01"]))/sd(SAMPresults[,"gam01"])
  CONDresults[counterPOP,"rb0gam02"]<-(mean(SAMPresults[,"se0gam02"])-
sd(SAMPresults[,"gam02"]))/sd(SAMPresults[,"gam02"])
  CONDresults[counterPOP,"rb0gam10"]<-(mean(SAMPresults[,"se0gam10"])-
sd(SAMPresults[,"gam10"]))/sd(SAMPresults[,"gam10"])
  CONDresults[counterPOP,"rbFPgam01"]<-(mean(SAMPresults[,"seFPgam01"])-
sd(SAMPresults[,"gam01"]))/sd(SAMPresults[,"gam01"])
  CONDresults[counterPOP,"rbFPgam02"]<-(mean(SAMPresults[,"seFPgam02"])-
sd(SAMPresults[,"gam02"]))/sd(SAMPresults[,"gam02"])
  CONDresults[counterPOP, "rbFPgam10"]<-(mean(SAMPresults[,"seFPgam10"])-
sd(SAMPresults[,"gam10"]))/sd(SAMPresults[,"gam10"])
  CONDresults[counterPOP, "rbBOOTgam01"]<-
(mean(SAMPresults[,"seBOOTgam01"])-
sd(SAMPresults[,"gam01"]))/sd(SAMPresults[,"gam01"])
  CONDresults[counterPOP, "rbBOOTgam02"]<-
(mean(SAMPresults[,"seBOOTgam02"])-
sd(SAMPresults[,"gam02"]))/sd(SAMPresults[,"gam02"])
  CONDresults[counterPOP, "rbBOOTgam10"]<-
(mean(SAMPresults[,"seBOOTgam10"])-
sd(SAMPresults[,"gam10"]))/sd(SAMPresults[,"gam10"])
```

```
 CONDresults[counterPOP, "mse0gam01"]<-mean(((SAMPresults[,"se0gam01"])-
sd(SAMPresults[,"gam01"]))^2)
 CONDresults[counterPOP, "mse0gam02"]<-mean(((SAMPresults[,"se0gam02"])-
sd(SAMPresults[,"gam02"]))^2)
 CONDresults[counterPOP, "mse0gam10"]<-mean(((SAMPresults[,"se0gam10"])-
sd(SAMPresults[,"gam10"]))^2)
 CONDresults[counterPOP, "mseFPgam01"]<-mean(((SAMPresults[,"seFPgam01"])-
sd(SAMPresults[,"gam01"]))^2)
 CONDresults[counterPOP, "mseFPgam02"]<-mean(((SAMPresults[,"seFPgam02"])-
sd(SAMPresults[,"gam02"]))^2)
 CONDresults[counterPOP, "mseFPgam10"]<-mean(((SAMPresults[,"seFPgam10"])-
sd(SAMPresults[,"gam10"]))^2)
 CONDresults[counterPOP, "mseBOOTgam01"]<-
mean(((SAMPresults[,"seBOOTgam01"])-sd(SAMPresults[,"gam01"]))^2)
 CONDresults[counterPOP, "mseBOOTgam02"]<-
mean(((SAMPresults[,"seBOOTgam02"])-sd(SAMPresults[,"gam02"]))^2)
 CONDresults[counterPOP, "mseBOOTgam10"]<-
mean(((SAMPresults[,"seBOOTgam10"])-sd(SAMPresults[,"gam10"]))^2)

 CONDresults[counterPOP, "Rmse0gam01"]<-
sqrt(CONDresults[counterPOP,"mse0gam01"])
 CONDresults[counterPOP, "Rmse0gam02"]<-
sqrt(CONDresults[counterPOP,"mse0gam02"])
 CONDresults[counterPOP, "Rmse0gam10"]<-
sqrt(CONDresults[counterPOP,"mse0gam10"])
 CONDresults[counterPOP, "RmseFPgam01"]<-
sqrt(CONDresults[counterPOP,"mseFPgam01"])
 CONDresults[counterPOP, "RmseFPgam02"]<-
sqrt(CONDresults[counterPOP,"mseFPgam02"])
 CONDresults[counterPOP, "RmseFPgam10"]<-
sqrt(CONDresults[counterPOP,"mseFPgam10"])
 CONDresults[counterPOP, "RmseBOOTgam01"]<-
sqrt(CONDresults[counterPOP,"mseBOOTgam01"])
 CONDresults[counterPOP, "RmseBOOTgam02"]<-
sqrt(CONDresults[counterPOP,"mseBOOTgam02"])
 CONDresults[counterPOP, "RmseBOOTgam10"]<-
sqrt(CONDresults[counterPOP,"mseBOOTgam10"])

 CONDresults[counterPOP, "ZEROcover0gam02"]<-
mean(SAMPresults[,"ZEROcover0gam02"])
 CONDresults[counterPOP, "ZEROcoverFPgam02"]<-
mean(SAMPresults[,"ZEROcoverFPgam02"])
 CONDresults[counterPOP, "ZEROcoverBOOTgam02"]<-
mean(SAMPresults[,"ZEROcoverBOOTgam02"])
 CONDresults[counterPOP, "PARAMcover0gam02"]<-
mean(SAMPresults[,"PARAMcover0gam02"])
```

```
  CONDresults[counterPOP, "PARAMcoverFPgam02"]<-
mean(SAMPresults[,"PARAMcoverFPgam02"])
  CONDresults[counterPOP, "PARAMcoverBOOTgam02"]<-
mean(SAMPresults[,"PARAMcoverBOOTgam02"])

}#POPreps loop END

CONDresults[counterPOP, "popID"]<-pop_id

#toc()#end timer

Save.file<-paste0("/work/edpsyc/ssvoboda/Cj60nj150/results.",pop_id,".csv")

write.table(CONDresults, file = Save.file, quote = FALSE, sep = ",", row.names =
FALSE, col.names = TRUE)

#print(CONDresults)
```

APPENDIX C: CONTINUOUS PREDICTORS' RESULTS FROM THE BINARY PREDICTOR STUDY

Figure C.1. Percentage relative bias in *SE*s for γ_{01} in the binary predictor study.

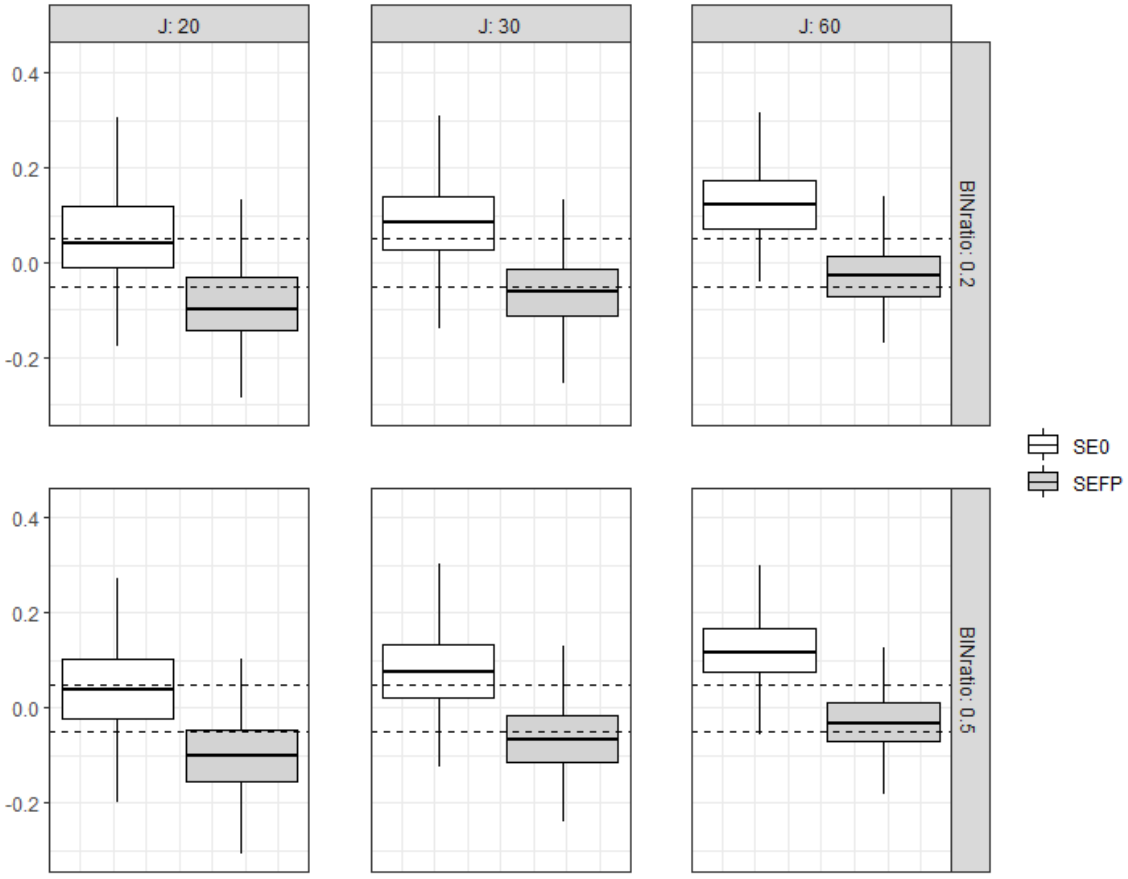


Table C.1. *Mean Square Error and Root Mean Square Error for* γ_{01} *in Binary Predictor Study*

	MSE (RMSE)	
J	SE_0	SE^{FP}
20	.0078 (.0869)	.0071 (.0830)
30	.0034 (.0576)	.0027 (.0511)
60	.0011 (.0321)	.0006 (.0234)

Note. MSE = mean square error; RMSE = root mean square error; J = number of clusters in sample; SE_0 = unadjusted standard error estimates; SE^{FP} = FPC adjustment estimates. Values were rounded to the nearest .0001.

Figure C.2. Percentage relative bias in *SE*s for γ_{10} in the binary predictor study.

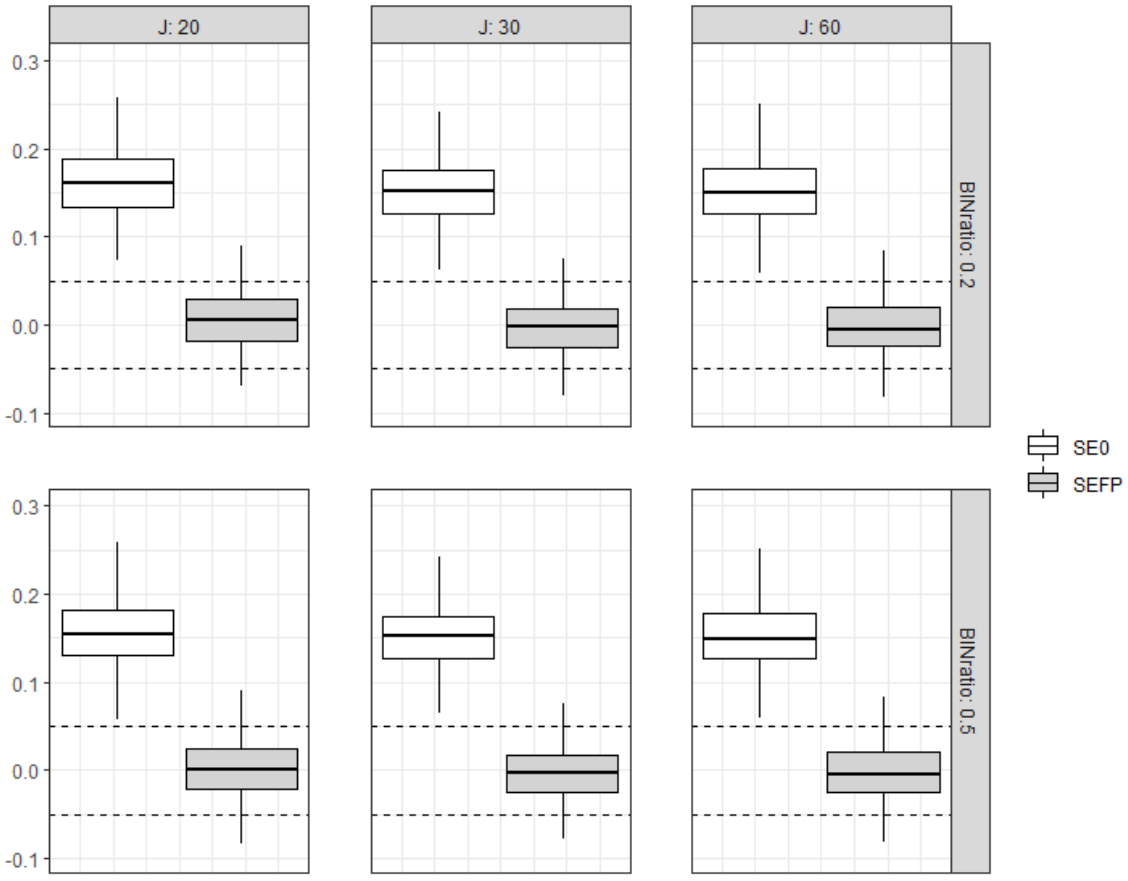


Table C.2. *Mean Square Error and Root Mean Square Error for* γ_{10} *in Binary Predictor Study*

	MSE (RMSE)	
J	SE_0	SE^{FP}
20	.0013 (.0352)	.0005 (.0225)
30	.0007 (.0257)	.0002 (.0151)
60	.0003 (.0160)	<.0001 (.0078)

Note. MSE = mean square error; RMSE = root mean square error; J = number of clusters in sample; SE_0 = unadjusted standard error estimates; SE^{FP} = FPC adjustment estimates. Values were rounded to the nearest .0001.